Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estudo comparativo de Abordagens Semi-Supervisionadas para Análise de Sentimentos em Tweets

Maria Fernanda do Carmo, Rodrigo Doria Vilaça

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Orientador
Prof. Dr. Vinicius Ruela Pereira Borges

Brasília
2020

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Estudo comparativo de Abordagens Semi-Supervisionadas para Análise de Sentimentos em Tweets

Maria Fernanda do Carmo, Rodrigo Doria Vilaça

Monografia apresentada como requisito parcial
para conclusão do Bacharelado em Ciência da Computação

Prof. Dr. Vinicius Ruela Pereira Borges (Orientador)
CIC/UnB

Prof.a Dr.a Roberta Barbosa Oliveira
CIC/UnB

Prof. Dr. Jan Mendonca Correa
CIC/UnB

Prof. Dr. Marcelo Grandi Mandelli
Coordenador do Bacharelado em Ciência da Computação

Brasília, 27 de Novembro de 2020

Dedicatória

Dedicamos esse trabalho à todos nossos colegas que foram espertos o suficiente para desistir.

Agradecimentos

Gostaríamos de agradecer ao nosso orientador Professor Doutor Vinícius Borges, que auxiliou na construção da nossa pesquisa e se disponibilizou para ajudar sempre que necessário.
Agradecemos aos amigos que nos ajudaram na realização deste projeto rotulando centenas de Tweets, e também todos que ajudaram direta e indiretamente em todo o processo.
Maria Fernanda: Agradeço à minha família e aos meus amigos por todo o apoio, não teria chegado aqui sem vocês. Especialmente minha irmã por nunca ter deixado eu me acomodar. E à tia Patrícia pelo melhor exemplo de mulher que eu já tive, você teria adorado isso. Por fim e mais importante agradeço ao Rodrigo, muito obrigada por toda a paciência e loucura. E por incentivar minhas piores ideias. Obrigada por não me deixar fazer isso sozinha.
Rodrigo: Agradeço a minha noiva, que estava lá para mim em todos os ótimos, bons, maus e péssimos momentos desta trajetória, sempre me mantendo motivado e esperançoso, confiando em meu potencial como ninguém. Agradeço a minha família por me apoiar durante toda minha graduação. Por fim, agradeço a minha amiga Maria Fernanda, que provavelmente será a única pessoa a saber realmente todos os perrengues que tivemos que passar para chegar até aqui, e por revezar os surtos comigo durante todo o processo.

Resumo

Existem várias abordagens para desenvolver métodos de aprendizado de máquina voltados para análise de sentimentos. Há uma carência, no entanto, de estudos e conjuntos de dados usando *tweets* na língua portuguesa para análise de sentimentos. Adicionalmente, visto a dificuldade de se encontrar conjuntos de dados rotulados para a implementação, abordagens semi-supervisionadas podem ser uma alternativa para contornar este problema, podendo-se usar conjuntos de dados com apenas uma parte dos dados rotulados.

Este trabalho faz uma comparação de diferentes métodos de aprendizado de máquina semi-supervisionados em relação à métodos supervisionados, aplicados à análise de sentimentos para, entre outros fins, detecção e classificação de polaridades de textos, e suas variadas formas de implementação e análise. Para esse propósito, uma metodologia é proposta para a classificação de sentimentos em *tweets* utilizando dois conjuntos de dados, sendo um criado inteiramente de *tweets* na língua Portuguesa, relacionados à Universidade de Brasília, e também um conjunto de *tweets* em língua Inglesa. Os *tweets* foram rotulados em positivo, negativo ou neutro, à fim de que se possa utilizar métodos de aprendizado de máquina supervisionados e semi-supervisionados. Basicamente o método consiste nas etapas de pré-processamento dos dados, extração de características e classificação utilizando os modelos *Support Vector Machines* (SVM), *Naive Bayes*, *Label Propagation* e *k-Nearest Neighbors* (KNN). Por fim, a performance dos classificadores é avaliada utilizando a F1-Score, levando às conclusões em relação à eficácia do aprendizado semi-supervisionado comparado ao supervisionado, afim de entender melhor como a abordagem semi-supervisionada se comporta neste cenário.

Palavras-chave: Análise de Sentimentos, Twitter, Aprendizado de Máquina, Aprendizado Semi-Supervisionado

Abstract

There are several approaches that consider machine learning methods in the sentiment analysis field. However, there is a lack of studies and datasets in Portuguese in this scope. Additionally, due to the complexity to find labeled datasets for the studies, semi-supervised approaches can be an alternative to study this problem, making it possible to employ datasets with only part of labeled data.

This work makes a comparison of different semi-supervised machine learning methods in relation to supervised methods, for sentiment analysis tasks. Specifically, the key idea is to detect and classify tweets according to predefined polarities, as well as, analyzing their various forms of implementation and analysis. For this purpose, a methodology is proposed to classify sentiments in tweets using two corpora, in which one was created from Portuguese tweets, collected from profiles related to the University of Brasilia, while the other one is constituted by tweets in English language. The tweets were classified as positive, negative or neutral, by considering supervised and semi-supervised machine learning techniques. Respectively, data pre-processing, feature extraction and classification were performed using the models Support Vector Machines (SVM), Naive Bayes, Label Propagation and k-Nearest Neighbors (KNN). Finally, the classifier's performance is analysed using F1-Score, leading to conclusions regarding the effectiveness of semi-supervised learning compared to supervised learning, in order to better understand the behavior of semi-supervised approaches in this scenario.

Keywords: Sentiment Analysis, Twitter, Machine Learning, Semi-Supervised Learning

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

As redes sociais são um espaço para as pessoas dividirem suas opiniões e interagirem com pessoas que não necessariamente estão fisicamente próximas e que não precisam estar disponíveis no momento do envio de mensagens para poder recebê-las. Atualmente, as redes sociais são a forma mais popular de disseminar informações, pois possuem um grande alcance, rapidez e facilidade. Todas as redes sociais somam 3.8[1] bilhões de usuários, ou seja, cerca de metade do planeta já está conectada em alguma rede social.

O Twitter[2] é uma rede social em que usuários podem postar pequenos textos expressando sua opinião acerca de diversos temas, com no máximo 280 caracteres. Estes pequenos textos são chamados de *tweets* e são importantes fontes de informação em pesquisas envolvendo as áreas empresarial e acadêmica. *Tweets* estão sendo utilizados para fazer previsões de crimes [1] e verificar a quantidade de *Fake News* em campanhas presidenciais nas eleições do Estados Unidos [2]. Sendo assim, *tweets* se tornaram uma valiosa fonte de informação devido à simplicidade de se obter grandes quantidades de dados textuais que são disponibilizados de forma pública.

Diariamente, mais de quinhentos milhões de *tweets* são postados[3], sendo considerado uma grande fonte de dados disponível para análise tanto de pessoas físicas como jurídicas. Apesar do Brasil ser o sexto país que mais usa o Twitter no mundo [4], há uma escassez de pesquisas relacionadas à análise de sentimentos utilizando dados na Língua Portuguesa [3]. Uma parte desse trabalho foi dedicada para classificar *tweets* conforme as polaridades

[1] https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media. Acessado em Setembro de 2020.

[2] `https://twitter.com/`

[3] blog.statusbrew.com/social-media-statistics-2018-for-business. Acessado em Setembro de 2020.

[4] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries. Acessado em Setembro de 2020.

positivo, negativo ou neutro que foram escritos em Português, obter um conjunto de dados com os *tweets* rotulados e comparar o desempenho de alguns modelos de Inteligência Artificial na classificação desses *tweets*.

Conforme Silva et al. [4] existem três formas principais de se fazer análise de sentimentos: por aprendizado de máquina, por abordagens baseadas em léxicos, ou ainda, híbrida. O Aprendizado de Máquina é baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana e se divide em quatro subcategorias: supervisionado, em que todos os rótulos estão disponíveis e são utilizados na análise, semi-supervisionado que contém uma parcela dos dados rotulados e uma parcela não rotulada, não-supervisionado em que os dados de interesse não possuem informações de rótulos ou categorias, e reforço onde um programa deve desempenhar um determinado objetivo e são fornecidos *feedbacks* durante a execução, reforçando o aprendizado. Geralmente, abordagens léxicas fazem uso de dicionários contendo as palavras da língua escolhida e sua polaridade entre outras informações relevantes, enquanto que abordagens baseadas em contexto dependem de análises estatísticas e semânticas para encontrar a polaridade do texto.

Neste trabalho, serão utilizadas as abordagens de aprendizado de máquina supervisionado e semi-supervisionado, com objetivo de realizar um estudo comparativo em análise de sentimentos em *tweets*. Devido às dificuldades de se encontrar conjuntos de textos na língua Portuguesa rotulados e à complexidade de se rotular um conjunto de dados, decidiu-se estudar abordagens de análise de sentimentos baseadas em aprendizado semi-supervisionado. Nesse sentido, a ideia é verificar se podem-se obter desempenhos de classificação satisfatórios utilizando uma menor quantidade de dados rotulados em relação à abordagem supervisionada. Assim, caso tal desempenho seja favorável em relação às abordagens semi-supervisionadas, pode-se poupar o esforço, às vezes manual, de ter que rotular grandes quantidade de dados.

A principal contribuição deste trabalho consistiu no desenvolvimento de um estudo comparativo de abordagens semi-supervisionadas para a classificação de polaridade em *tweets*, sabendo-se da dificuldade de anotar conjuntos de dados para tarefas de aprendizado de máquina. Embora a literatura apresente abordagens não-supervisionadas e semi-supervisionadas para esse propósito, poucas pesquisam consideraram corpus em língua Portuguesa em seus experimentos, como o trabalho de Aguiar et al. [3]. Assim, outra importante contribuição foi a criação de um corpus rotulado de *tweets* em língua Portuguesa relacionados com a Universidade de Brasília (UnB) e que podem ser úteis para encontrar conhecimento referente às opiniões de pessoas relacionadas à UnB. A escolha de *tweets* relacionados à UnB foi motivada por uma preocupação crescente com a saúde mental do corpo discente, docente e funcionários dentro do ambiente universitá-

rio. Além disso, os *tweets* relacionados com a UnB são uma fonte interessante de análise. Por exemplo, existem *tweets* que indicam sentimentos positivos quando os alunos ingressam na universidade, e também *tweets* negativos referentes aos métodos de avaliação das disciplinas.

1.2 Objetivo

Este trabalho possui como objetivo investigar a realização de experimentos de abordagens semi-supervisionadas para análise de sentimentos nas línguas Inglesa e Portuguesa por meio dos modelos de classificação *Support Vector Machines (SVM)*, *Naive Bayes*, *K-nearest-neighbor (KNN)* e *Label Propagation*. Para alcançar este objetivo, foram definidos os seguintes objetivos específicos:

- Realização de estudo de trabalhos relacionados à análise de sentimentos e levantamento de fundamentos teóricos;
- Construção de um novo conjunto de dados de *tweets* em Português, relacionados à Universidade de Brasília;
- Exploração de estratégias de pré-processamento e caracterização dos *tweets* nas línguas Inglesa e Portuguesa para viabilizar o emprego dos modelos de aprendizado de máquina;
- Realização do treinamento dos modelos de classificação em abordagens supervisionada e semi-supervisionada;
- Realização de experimentos nos modelos de classificação treinados e análise de seus resultados.

1.3 Estrutura do Documento

Este trabalho é composto pelos seguintes capítulos:

- Capítulo 2: Introduz conceitos de Mineração de Textos e Aprendizado de Máquina, apresentando a teoria dos modelos de classificação utilizados, assim como as técnicas de avaliação de performance da classificação.
- Capítulo 3: Realiza a revisão de literatura, em que foram selecionados alguns artigos que exploram o tema de análise de sentimentos utilizando diversos modelos e técnicas de classificação diferentes.

- Capitulo 4: Descreve os métodos propostos baseados em aprendizado de máquina, desde a coleta de dados até a forma de utilização do aprendizado de máquina através dos modelos de classificação escolhidos.
- Capítulo 5: Apresenta os resultados obtidos, utilizando as técnicas de avaliação de performance explicitadas.
- Capítulo 6: Discute as conclusões obtidas a partir dos resultados dos experimentos e apresenta sugestões para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo descreve os fundamentos teóricos relacionados com os métodos mais empregados na área de mineração de textos e análise de sentimentos. Inicialmente, o capítulo introduz os fundamentos sobre textos e as técnicas usadas para o pré-processamento e caracterização. Na parte de Aprendizado de Máquina, são citados quais modelos serão usados, as particularidades de cada um e como foi feita a avaliação de performance dos classificadores.

2.1 Textos

Existem diversos tipos de textos disponíveis na internet: livros, documentos, receitas, notas fiscais, críticas de filmes, mensagens, entre outros. Textos podem ser divididos em dados estruturados, semi-estruturados e não estruturados. Dados estruturados seguem uma estrutura previamente estipulada e que não pode ser alterada, são o caso de bancos de dados e formulários. Dados semi-estruturados têm uma estrutura mais flexível que em geral pode sofrer alterações, por exemplo arquivos JSON e XML. Dados não estruturados são flexíveis e dinâmicos. Estima-se que mais de 85% de todas as informações de negócios são dados não-estruturados provindos de artigos, e-mails, material de marketing, redes sociais, entre outras [5].

Um *tweet* é uma mensagem publicada no Twitter, pode conter texto com um máximo de 280 caracteres, fotos, GIF e/ou vídeo. *Tweets* aparecem na página de perfil do remetente e na linha do tempo de qualquer usuário que siga o remetente. No caso de perfis abertos, os *tweets* podem aparecer nas buscas por palavras chaves[1]. O excesso de *tweets* também pode ser desfavorável, uma vez que cerca de 200 bilhões de *tweets* são publica-

[1]https://help.twitter.com/pt/using-twitter/types-of-tweets. Acessado em Setembro de 2020.

dos por ano, sendo necessário o uso de técnicas de mineração de textos para analisar e processar esse grande volume de dados[2].

2.2 Mineração de textos

A mineração de textos tem como principal objetivo fazer uso do acesso à informação para ajudar usuários a analisar dados textuais facilitando assim a tomada de decisões. Geralmente, essa análise procura descobrir padrões interessantes como tendências e pontos dissonantes, e tem como principal dificuldade a alta dimensionalidade resultante do processo de extração de características de textos em linguagem natural [6].

2.2.1 Pré-processamento de textos

Remoção de Stop words

Stop words são as palavras que agregam pouco valor semântico às frases, sendo geralmente compostas por artigos, preposições, pontuação, conjunções e pronomes. Wives e Lo [7] consideram *stop words* como palavras vazias, pois aparecem na maioria dos textos e não acrescentam informações relevantes para suas polaridades. Usualmente são feitas listas de *stop words*, as *stop lists* que são retiradas dos textos na etapa de pré-processamento para diminuir o ruído e aumentar a acurácia em tarefas de mineração de textos. As *stop words* na língua Portuguesa costumam incluir “o”, “a”, “do”, “de”, enquanto que em Língua Inglesa compreendem palavras como “the”, “is”,“and”.

A remoção de *stop words* de textos pode afetar a performance nas tarefas de classificação. Saif et al. [8] concluíram que retirar *stop words* genéricas pode ter um impacto negativo na acurácia do modelo. No entanto, remover as palavras que aparecem apenas uma vez no documento é o melhor custo benefício entre diminuir a dimensão dos dados e aumentar a acurácia.

Stemização

Palavras em geral são compostas por prefixo, sufixo e radical. No radical está o verdadeiro significado de cada palavra, sendo fundamental para a análise de sentimentos. Prefixos não costumam ser dispensados, pois podem alterar a polaridade das palavras. Por exemplo: A palavra “desejável”, que tem polaridade positiva e “indesejável”, com polaridade negativa. Comumente, sufixos não alteram o significado das palavras e podem ser retirados.

A stemização é o processo que retira o final das palavras com o objetivo de reduzir as diferentes terminações que possuem o significado similar para uma única palavra.

[2]https://www.internetlivestats.com/twitter-statistics. Acessado em Setembro de 2020.

Geralmente, esse processo é implementado de forma heurística [9]. Nas palavras "democracia", "democrático" e "democratizar" a stemização provavelmente diminuiria todas para "democra" ou "democr".

Lematização

A lematização é um processo com o objetivo análogo ao da stemização, porém é implementado de maneira distinta. Ao invés de retirar heuristicamente o final das palavras, na lematização considera-se a análise morfológica e o vocábulo, sendo que a intenção é reduzir a palavra ao seu lema de preferência fazendo consultas a um dicionário. Exemplificando, a palavra "melhor" tem "bom" como lema, enquanto que na stemização isso seria ignorado. Por causa da sua complexidade, a lematização é mais comum em trabalhos de processamento de linguagem natural [9].

Remoção de Caracteres

Na literatura, é muito comum se fazer a remoção de alguns caracteres específicos, porém o caractere a ser removido depende muito do conjunto de textos. Alguns trabalhos removem pontuação porque consideram uma parte irrelevante na análise [10]. Em outros casos, é fundamental não retirar a pontuação, como em projetos que fazem análise de sentimentos baseada em *emoticons* [11]. No conjunto de dados desse projeto, fez-se necessária a retirada das URLs dos *tweets*.

2.2.2 Caracterização de textos

Bag of Words

Bag of Words (BoW), traduzido da língua inglesa como saco de palavras, é uma forma de representar termos em vetores. Qualquer informação relativa à ordem ou estrutura de palavras em um documento é descartada e são mantidas somente o conjunto de termos usados e sua frequência. Cada palavra do texto será considerada como um atributo [12]. Podemos demonstrar a aplicação no seguinte poema de Nikita Gill:
Let it hurt
Let it bleed
Let it heal
And let it go
O poema contém 13 palavras, porém apenas 7 distintas:
('let', 4), ('it', 4), ('hurt', 1), ('bleed', 1), ('heal', 1), ('and', 1), ('go', 1).

Poema	Let	it	hurt	bleed	heal	And	go
Let it hurt	1	1	1	0	0	0	0
Let it bleed	1	1	0	1	0	0	0
Let it heal	1	1	0	0	1	0	0
And let it go	1	1	0	0	0	1	1

Tabela 2.1: Tabela de Distribuição do Poema usando Bag of Words.

Cada frase será representada por um vetor de zeros e 1's, indicando respectivamente a ausência e presença de determinado vocábulo do dicionário na frase considerada, conforme pode-se observar na Tabela 2.1.

Observa-se que quanto maior é o vocabulário do documento, mais esparso será o vetor, sendo uma das desvantagens do uso de *BoW*, pois o tamanho do texto influencia na quantidade de memória e no uso dos recursos computacionais exigidos por esse método. Outra limitação desta técnica é não considerar o significado semântico de cada frase, uma vez que a alteração da ordem das palavras altera a semântica da frase, mas não altera a representação BoW em relação à frase original.

Term Frequency - Inverse Document Frequency

Term Frequency - Inverse Document Frequency (TF-IDF), é uma estatística numérica obtida pela multiplicação da frequência do termo pela frequência inversa dos documentos. A frequência dos termos é um indicador que mostra a quantidade de vezes que um mesmo termo aparece em um documento, e para evitar comparações desproporcionais, devido ao tamanho de cada documento, é calculada pelo número de vezes que uma palavra apareceu dividida pela quantidade total de palavras do documento [13]. A frequência inversa dos documentos atribui um peso maior para as palavras que são menos mencionadas no documento, é um método que busca equilibrar a frequência do termo, diminuindo a importância de palavras que costumam aparecer muitas vezes em textos, por exemplo artigos e pronomes, e agregar maior peso a palavras incomuns no documento mas que provavelmente têm maior relevância semântica.

Quaiser e Ali [13] apontam a incapacidade do algoritmo de reconhecer palavras parecidas como a maior limitação do TF-IDF, por exemplo "ano" e "anos" seriam palavras contadas como diferentes e independentes o que pode ocasionar resultados inesperados, por essa razão é recomendado o uso do TF-IDF com outros métodos que possuam em seu processo uma abordagem semântica, por exemplo stemização.

2.3 Aprendizado de Máquina

O aprendizado de máquina foi definido por Tom M. Mitchell [14] da seguinte forma: "Diz-se que um programa de computador aprende pela experiência E, com respeito a algum tipo de tarefa T e performance P, se sua performance P nas tarefas em T, na forma medida por P, melhoram com a experiência E". Pode-se verificar que as tarefas envolvidas se tornam importantes para este conceito. Tais tarefas de aprendizado de máquina podem ser divididas em quatro categorias de acordo com o *feedback* disponível para um sistema de aprendizado: o aprendizado supervisionado, o aprendizado não-supervisionado, o aprendizado semi-supervisionado e o aprendizado por reforço. Apenas as três primeiras categorias serão brevemente explicadas, uma vez que o aprendizado por reforço não será abordado neste trabalho.

2.3.1 Aprendizado supervisionado

No aprendizado supervisionado, é fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido [15]. Em outras palavras, um algoritmo de aprendizado aprende os padrões dos exemplos de entrada com base nas suas respectivas categorias, com objetivo de aprender uma regra geral que mapeia as entradas para as saídas.

A maioria dos estudos sobre análise de sentimentos em *tweets* utilizam algoritmos de aprendizado supervisionado para produzir modelos de classificação de sentimentos, e tais algoritmos necessitam de um conjunto de treinamento formado por dados rotulados, onde os rótulos são as classes (e.g., positivo, neutro e negativo) para cada *tweet* [4].

2.3.2 Aprendizado Não-Supervisionado

Já no aprendizado não-supervisionado, o indutor analisa os exemplos fornecidos e tenta determinar se alguns deles podem ser agrupados conforme algum critério de similaridade, formando agrupamentos ou *clusters*. Após determinar os agrupamentos, normalmente, é necessária uma análise para determinar o significado de cada agrupamento no contexto do problema sendo estudado [15]. Por isso, é o tipo de abordagem mais comum na existência de dados não-rotulados, onde o algoritmo de aprendizado necessita extrair estruturas das entradas fornecidas, podendo descobrir novos padrões, ou sendo um meio para atingir tal fim.

2.3.3 Aprendizado Semi-Supervisionado

Por fim, no aprendizado semi-supervisionado, são utilizados tanto dados rotulados quanto dados não rotulados [16]. Essa abordagem é apropriada para ser estudada nesta pesquisa, considerando a dificuldade de se encontrar coleções de *tweets* em língua Portuguesa rotuladas para tarefas de análise de sentimentos. Desta forma, é possível treinar um modelo de classificação utilizando uma quantidade maior de informação, e com a utilização de dados rotulados e não rotulados pode-se obter um classificador que possui eficácia superior em relação a um classificador treinado utilizando todos os dados rotulados [17].

Silva et al. [4] identificam três categorias de abordagens semi-supervisionada para análise de sentimentos em *tweets*:

- (i) *graph-based methods* (ou métodos baseados em grafos),
- (ii) *wrapper-based methods* (e.g., auto-treinamento e co-treinamento), e
- (iii) *topic-based methods*.

Neste trabalho, foram utilizados modelos de classificação que se encaixam nas categorias (i) e (ii), por serem mais empregadas nos trabalhos relacionados estudados. A terceira categoria não foi explorada neste projeto, porém há planos de abordá-la em projetos futuros.

Nos métodos baseados em grafos, os rótulos dos dados de entrada são propagados para os dados não rotulados. O processo de propagação de rótulos necessita do cálculo das similaridades entre as instâncias dos dados [4]. Já *wrapper-based methods* utilizam algoritmos de aprendizado supervisionado de maneira iterativa. Em cada iteração, uma certa quantidade de instâncias não rotuladas é rotulada pela função de decisão que é aprendida e incorporada ao conjunto de treinamento [4]. Neste projeto, a técnica utilizada para representar este método foi o auto-treinamento, em que "o processo que está aprendendo utiliza suas próprias predições para se ensinar" [17].

2.3.4 Modelos de classificação

Foram utilizados quatro modelos de classificação neste trabalho, que serão descritos nas seções a seguir. Nas ramificações abaixo, considera-se um conjunto de dados composto por n instâncias $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}\}$, em que $x_i = (x_{i,1}, x_{i,2}, ...x_{i,m})$.

Máquina de Vetores de Suporte (SVM)

Máquina de Vetores de Suporte, ou *Support Vector Machine* (SVM), é um modelo de classificação que analisa dados reconhecendo padrões e pode ser utilizado para problemas de

classificação ou regressão, sendo mais comumente usado para problemas de classificação. O algoritmo foi elaborado com base na teoria de aprendizagem estatística, ou teoria VC, desenvolvida por Vapnik e Chervonenkis desde os anos 70, caracterizando propriedades de máquinas de aprendizados fazendo com que elas possam generalizar bem dados não conhecidos [18].

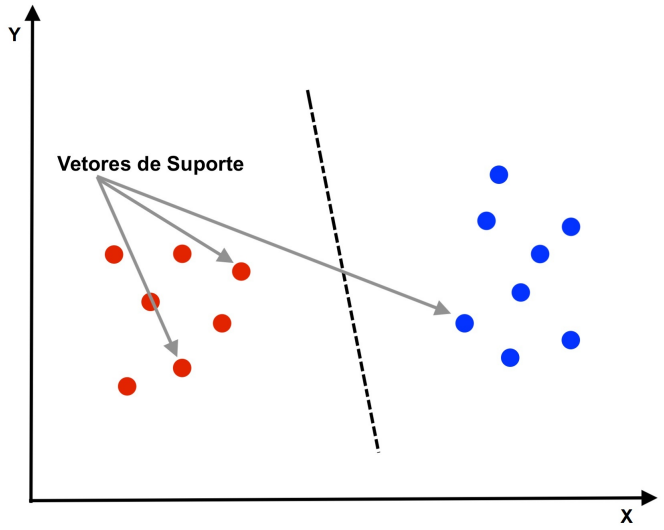

Figura 2.1: Exemplificação de Vetores de Suporte.

Dado um conjunto de exemplos de treinamento, em que cada um pertence a uma categoria, o algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos à uma categoria específica. O objetivo do SVM é encontrar o melhor hiperplano que divide tais categorias, através dos vetores suporte, definidos como padrões críticos que sozinhos determinam o hiperplano ótimo [19]. Pode-se ficar mais claro observando a Figura 2.1 em que, os eixos X e Y representam a área onde exemplos de treinamento estão localizados, identificados por círculos vermelhos e azuis que representam as duas classes, e a linha que segrega claramente as duas categorias seria o hiperplano gerado pelo algoritmo SVM.

Naive Bayes

O algoritmo *Naive Bayes*, é um modelo de classificação probabilístico baseado no teorema de Bayes. A partir de exemplos de treinamento para criar um modelo probabilístico baseado nas características dos dados, é calculada a probabilidade de um evento ocorrer (e.g. um dado ser classificado em uma determinada classe) dado que outro evento já ocorreu. Este modelo atribui a classe mais provável a um dado exemplo descrito pelo seu vetor de características [20].

O algoritmo possui o termo *Naive*, ou ingênuo, em seu nome por assumir que todos os atributos da amostra são independentes entre si, dado o contexto da classe [21]. Em outras palavras, a presença de uma determinada característica nos dados não tem nenhuma relação com as outras para o cálculo da probabilidade de determinado evento ocorrer. Sabendo disto, considerando uma classe C, estes classificadores podem ser simplificados, se resumindo à seguinte fórmula:

$$P(\mathbf{X}|C) = \prod_{i=1}^{n} P(X_i|C) \tag{2.1}$$

Apesar de possuir tal suposição irrealista, o classificador resultante desta prática, conhecido por *Naive Bayes*, é bem-sucedido na prática [20].

KNN

K-nearest neighbors (KNN) é um dos mais fundamentais métodos de classificação e apresenta bons resultados em estudos com pouco ou nenhum conhecimento prévio sobre o conjunto de dados, por ser um algoritmo de aprendizado não-paramétrico. O KNN foi introduzido nos anos 50 em um artigo não publicado de Fix e Hodges como um método não-paramétrico para classificação de padrões que ficou conhecido como a regra do *k-Nearest Neighbors* [22].

No algoritmo KNN, K é o número de vizinhos mais próximos e o fator decisivo mais importante. O algoritmo consiste basicamente de três passos[3]:

1. Calcular distância entre x e os outros pontos (y), usando mais comumente, a distância Euclidiana, dada pela seguinte fórmula:

$$d(x, y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \tag{2.2}$$

2. Identificar os K vizinhos mais próximos.
3. Cada vizinho vota em sua classe, e a classe com mais votos se torna a classificação de x.

Estes passos podem ser visualmente exemplificados pela Figura 2.2.

[3] `https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn`. Acessado em Setembro de 2020.

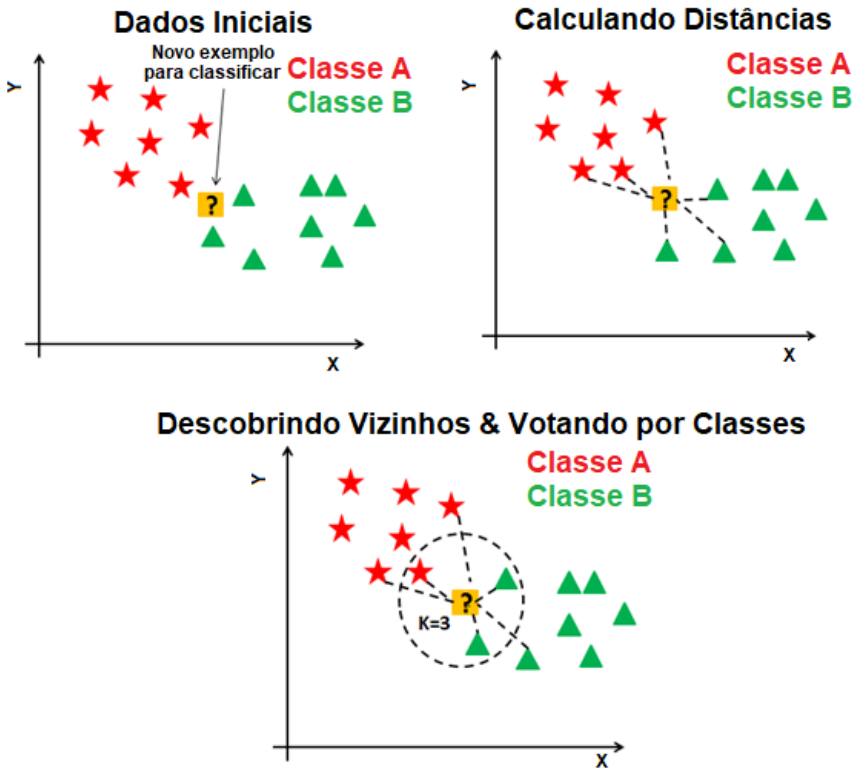

Figura 2.2: Funcionamento do KNN.[3]

Label Propagation

Label Propagation é um algoritmo de aprendizado de máquina semi-supervisionado que atribui rótulos (*labels*) para dados não rotulados. A partir de um pequeno conjunto de dados rotulados, inicia-se um processo de propagação dos rótulos para os dados não-rotulados [23].

Segundo Johson et al. [24], a ideia por trás do *Label Propagation* é construir um grafo com pesos $G = (V, E, W)$, em que V é o conjunto de vértices constituído de usuários, *tweets* e outras características e E é o conjunto de arestas conectando os vértices, e W é o conjunto de pesos associados às arestas, onde $W_{i,j}$ representa o peso da aresta (i, j). Com tal estrutura de grafo, a distribuição de rótulos é semeada em um conjunto inicial de vértices e depois espalhada através do grafo. Em cada iteração deste algoritmo, os rótulos dos vértices são atualizados para o rótulo mais comum entre os vizinhos mais próximos até atingir convergência, ou o número de iterações máximo[4], conforme pode ser visto na Figura 2.3. Casos de empate são decididos de forma randômica.

[4] `https://neo4j.com/blog/graph-algorithms-neo4j-label-propagation/`. Acessado em Setembro de 2020.

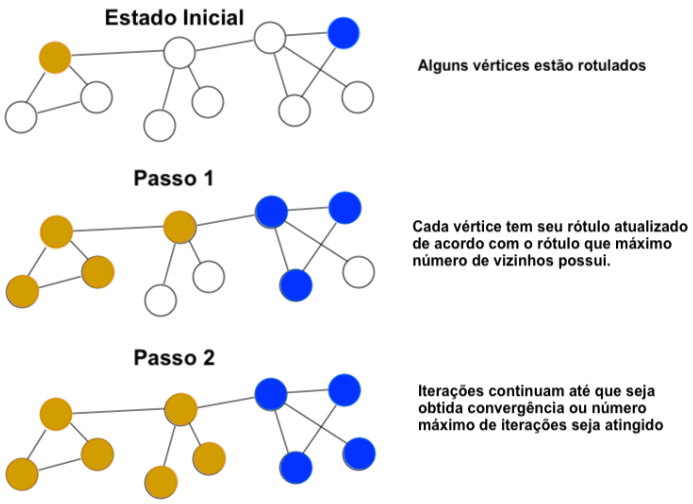

Figura 2.3: Funcionamento do *Label Propagation*[4].

2.3.5 Avaliação de Performance dos Classificadores

Existem diversas métricas para avaliar a acurácia dos classificadores. Na literatura relacionada à aprendizado de máquina, as métricas de precisão, revocação (i.e. *recall*) e *F1 Score* são as métricas mais populares, obtidas a partir da matriz de confusão.

Matriz de Confusão

A matriz de confusão apresenta a performance de uma classificação de um classificador à respeito de um conjunto de dados de teste. Em um problema de classificação binária, por exemplo, é uma matriz de duas dimensões, onde uma delas é a classe verdadeira de um objeto e a outra, a classe que o classificador atribuiu [25].

Em outras palavras, a matriz de confusão, também conhecida como matriz de erro, é uma tabela que contabiliza as frequências de classificação para cada classe do modelo, permitindo a visualização da performance de um dado algoritmo. Cada linha da matriz representa instâncias da classe prevista, enquanto que cada coluna representa instâncias da classe real (ou vice-versa).

Através da Figura 2.4 é possível entender visualmente a matriz de confusão, que apresenta verdadeiros positivos (vp), verdadeiros negativos (vn), falsos positivos (fp) e falsos negativos (fn).

Precisão

Precisão (Pr) é definida como a proporção de verdadeiros positivos e o número total de positivos estimados pelo modelo de classificação [26]. Desta forma, a precisão é calculada

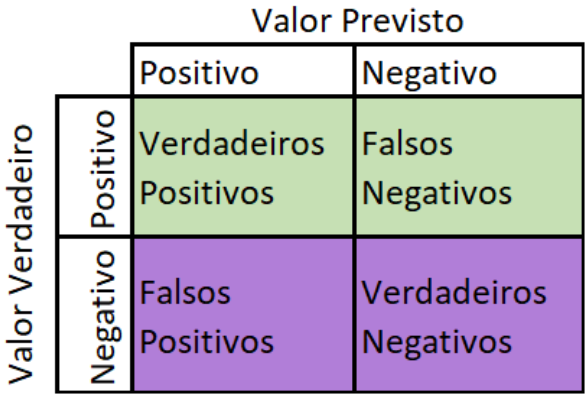

Figura 2.4: Exemplificação de Matriz de Confusão.

através da proporção:

$$Pr = \frac{vp}{(vp + fp)}. \tag{2.3}$$

A precisão é a habilidade do classificador de não rotular como positiva uma amostra que é negativa. Seu valor varia de 0 a 1, sendo 1 o melhor valor possível e zero o pior.

Revocação

Revocação (Re) é calculada através da proporção:

$$Re = \frac{vp}{(vp + fn)}. \tag{2.4}$$

A revocação é a habilidade do classificador de classificar corretamente todas as instâncias positivas. Seu valor varia de 0 a 1, sendo 1 o melhor valor possível e zero o pior.

F1 Score

O F1 Score, também conhecido como F Score balanceado ou *F-measure*, pode ser interpretado como uma média ponderada da precisão e da revocação. Esta métrica utiliza a precisão e a revocação para calcular a acurácia da previsão de acordo com a Eq. (2.5):

$$F1 = 2 * \frac{Pr * Re}{(Pr + Re)}, \tag{2.5}$$

em que F varia de 0 a 1, sendo 1 o melhor valor possível e zero o pior.

Os experimentos realizados neste trabalho foram possíveis graças ao entendimento dos fundamentos abordados neste capítulo.

Capítulo 3

Revisão de Literatura

A literatura possui vários trabalhos que exploram análise de sentimentos com diferentes conjuntos de dados. Neste capítulo, foram separados alguns desses trabalhos para resumir as técnicas utilizadas. A Seção 3.1 trata de artigos sobre a temática proposta neste trabalho e a Seção 3.2 explana a relevância deste trabalho.

3.1 Estudos em Análise de Sentimentos

Apesar de ser um tema relativamente novo, a literatura possui vários trabalhados que exploram análise de sentimentos com diferentes abordagens e conjuntos de dados. Não foram encontrados muitos trabalhos em que exploram este tema em conjuntos de dados em português, porém Aguiar et al. [3] usaram uma base de dados na língua portuguesa.

O artigo de Silva et al. [4] foi o que teve maior influência nessa monografia, dada a relevância do aprendizado de máquina semi-supervisionado e as comparações entre algoritmos feitas naquele projeto. A obtenção de dados rotulados costuma ser difícil e custosa, especialmente quando se trata de redes sociais. O aprendizado de máquina semi-supervisionado é uma alternativa interessante para esse problema, em que pode-se usar a grande quantidade de dados não rotulados disponíveis para complementar os dados rotulados no processo de treino dos algoritmos de análise de sentimentos. Foram utilizados seis conjuntos de dados disponibilizados pelo *International Workshop on Semantic Evaluation (SemEval)*. O conjunto de dados empregado para treino foi o SemEval 2013 com 11338 *tweets*, sendo 48% neutros, 15% negativos e 37% positivos. Dentre os conjuntos de dados usados para treino, três são oriundos do Twitter, Twitter2013, Twitter2014 e Tweeter Sarcasm 2014 compostos por respectivamente 1572, 982 e 33 *tweets* positivos, 601, 202 e 40 *tweets* negativos e 1640, 669 e 13 *tweets* neutros. Para conferir se a acurácia se mantêm caso os conjuntos de dados para treino tenham origem diferente da base de dados usada para teste, que é oriunda do Twitter, mais dois conjuntos de dados foram

usados. LiveJournal com 1142 textos, 37% positivos, 27% negativos e 36% neutros e um conjunto composto por mensagens de textos, SMS2013 constituído por 492 mensagens positivas, 304 mensagens negativas e 411 mensagens neutras.

Silva et al. [4] salientaram quatro pontos preocupantes com o estudo, sendo eles: apesar de ter uma limitação máxima de caracteres a média dos *tweets* é de 28 caracteres ocasionando matrizes muito esparsas quando se utiliza saco de palavras, *tweets* têm um vocabulário específico que frequentemente inclui palavras escritas errado, grande variedade de tópicos e dependência de *tweets* rotulados. Silva et al. [4] compararam algoritmos de aprendizado de máquina semi-supervisionado e um algoritmo de aprendizado de máquina não-supervisionado. Dentre as abordagens de aprendizado de máquina semi-supervisionado, co-treinamento obteve os melhores resultados para casos com menos *tweets* rotulados, seus melhores F-scores variando entre 65% e 80% com exceção do Twitter Sarcasm 2014 que apontou 50%. Destaca-se a indicação do auto-treinamento para conjuntos de dados com muitos *tweets* rotulados e para conjuntos de dados que contém sarcasmo, seus resultados variando entre 50% e 72%. O método não supervisionado obteve resultados piores em todos os casos, pois sua classificação consistia na análise de *tweets* com *hashtags* ou *emoticons*, correspondentes a apenas 842 *tweets* do conjunto de dados de treino, os outros 92.8% foram classificados como neutros, prejudicando a classificação dos *tweets*.

A proposta de Li et al. [27], é apresentar uma nova técnica para análise de sentimentos baseada em agrupamento, que possui vantagens comparadas às abordagens já existentes que são as técnicas simbólicas e métodos de aprendizado supervisionado. As técnicas simbólicas consistem em dar uma "pontuação de sentimento" para cada termo analisado. Uma abordagem baseada em aprendizado de máquina supervisionado empregou três classificadores que geram resultados de treinamento e eram testados usando *cross-validation*, obtendo acurácia entre 77.17% e 78.33%.

A pesquisa de Li et al. [27] considerou um conjunto de dados de 600 críticas, igualmente divididas em positivas e negativas, foram extraídas apenas palavras que eram adjetivos ou advérbios e então utilizaram o K-Means para agrupamento em 2 grupos. A metodologia empregada se baseia no algoritmo K-Means, em que foi aplicada a técnica TF-IDF nas críticas para melhorar a acurácia e um mecanismo de votação para que a divisão em dois grupos (concebidos como positivos e negativos). Por fim, a "pontuação de termos" da abordagem *symbolic techniques* também foi implementada para melhorar o resultado. O algoritmo supera problemas de baixa acurácia e instabilidade (que aconteciam antes da aplicação do TF-IDF e do mecanismo de votação no conjunto de dados para o agrupamento), não necessita de participação humana, sendo mais aplicável em situações reais do que os outros dois métodos.

Já em Aguiar et al. [3], foram usados *tweets* em português rotulados manualmente e disponibilizados pelo grupo de pesquisa MiningBr[1], que classifica os sentimentos entre positivo, negativo e neutro. A técnica usada foi o Comitê, que consiste na combinação da predição de algoritmos de classificação, sendo empregados os algoritmos *Naive Bayes*, SVM, Árvore de decisão, *Random Forest* e Regressão Logística. O peso de cada algoritmo no resultado do Comitê foi atribuído de acordo com a acurácia de cada um individualmente. O conjunto de *tweets* contém apenas 2516 registros, mas o Comitê teve melhor acurácia de 86.5%. Porém, no caso de um maior volume de dados, os autores citam o Naive Bayes como o mais interessante e ainda possui um tempo menor de execução. Também foram feitas diversas análises estatísticas e comparações com ferramentas existentes.

Hong et al. [28] apresentam o estudo de vários métodos de Modelagem em Tópicos, realizando experimentos qualitativos e quantitativos com mensagens de Twitter com duas tarefas diferentes: prever mensagens populares e classificar usuários e suas correspondentes mensagens em categorias comparando seus resultados. É usado um conjunto de dados de quase 2 milhões de mensagens e mais de 500 mil usuários. O artigo apenas explora esquemas que não necessitam de nenhuma modificação significativa aos modelos *LDA* ou *AT*. *Latent Dirichlet Allocation (LDA)* é um poderoso arcabouço para a modelagem de coleções de dados de contagem que recentemente tem sido aplicado a diversas tarefas, especialmente nas áreas de processamento de linguagem natural e visão computacional [29]. É possível que, utilizando extensões do LDA autores e mensagens sejam considerados simultaneamente, aumentando a utilidade e confiabilidade do dado. Para avaliar a primeira tarefa, são utilizadas precisão, revocação e *F-Score*, enquanto a segunda classificação considerou a acurácia. Concluiu-se que os tópicos obtidos por esquemas diferentes variam substancialmente e o *USER scheme* apresentou os melhores resultados. Para a primeira tarefa, conclui-se que empregar *TF-IDF* em conjunto com esquemas é uma boa prática e deve melhorar as precisões, e que *USER scheme* obtém resultados melhores em geral. Para a segunda tarefa, é inferido que, para textos pequenos, modelagens de tópicos ajudam nos modelos, e em textos mais longos é recomendado o uso de algo mais complexo.

O estudo de Dias e Becker [30] traz uma abordagem semi-supervisionada voltada para a detecção de posicionamento em *tweets* baseada em regras de sentimento. A detecção de posicionamento emprega a análise de sentimentos com o objetivo de identificar automaticamente se o autor de um texto é a favor, contra ou neutro em relação a um alvo, que pode ser uma entidade concreta (i.e. pessoa, organização ou local) ou uma entidade abstrata (i.e. uma causa ou afirmação). A abordagem de Dias e Becker [30] consiste em rotular automaticamente, através de regras, seis conjuntos de *tweets* de domínios distintos,

[1] `https://sites.google.com/site/miningbrgroup/home/publications`. Acessado em Setembro de 2020.

disponibilizados no SemEval2016, a fim de compor um corpus de treinamento para um método supervisionado de aprendizado, empregando o modelo de classificação SVM. A abordagem é semi-supervisionada, pois requer como entrada um conjunto de n-gramas que caracterizem apoio ou oposição à um determinado domínio. A seleção de tais n-gramas foi feita manualmente pelos autores, através de critérios subjetivos. Em seguida, a rotulação automática se dá através da aplicação de regras relacionadas aos n-gramas sobre os *tweets*. Os *tweets* nos quais alguma regra se aplica são rotulados automaticamente, enquanto os outros são descartados. A partir dos *tweets* rotulados automaticamente, é criado um *corpus* de treinamento utilizado para treinar um modelo de classificação supervisionado SVM. Um corpus de teste também é submetido à rotulação automática, e os *tweets* não rotulados nesta etapa são utilizados de entrada no classificador supervisionado gerado pela etapa anterior, obtendo-se os resultados com acurácias entre 14% e 75%, dependendo do domínio e de cada classe dentro dele (e.g. favorável, contrário e sem posicionamento). Em geral, o experimento retornou melhores resultados na detecção de posicionamentos contrários à um determinado domínio, devido ao fato de haver mais n-gramas de oposição nos domínios.

3.2 Considerações Finais

A partir dos trabalhos estudados, foi possível entender diversas técnicas usadas em análise de sentimentos, muitas das quais foram utilizadas nesta pesquisa, desde a coleta e pré-processamento de dados, até melhores modelos de classificação para cada caso e diversas maneiras para avaliações de performance. De toda forma, a maior parte dos trabalhos estudados explora a análise de sentimentos em abordagens supervisionadas usando conjuntos de dados em língua Inglesa, uma vez que não é uma tarefa fácil encontrar conjuntos de dados rotulados em língua Portuguesa.

Esta pesquisa pretende comparar abordagens semi-supervisionadas em conjuntos de dados em língua Inglesa e língua Portuguesa, este último sendo criado e rotulado manualmente por voluntários exclusivamente para essa pesquisa, sendo assim também uma contribuição para estudos futuros.

Capítulo 4

Metodologia Proposta

Este capítulo apresenta as etapas que compõem a metodologia adotada neste trabalho. Na Seção 4.1, é explicado como foi realizada a coleta de dados dos conjuntos de *tweets* considerados neste trabalho. Já na Seção 4.2, são abordadas as técnicas utilizadas para o pré-processamento dos dados. Na Seção 4.3, é explicada a técnica utilizada para converter os dados textuais em dados numéricos, ou seja, caracterização dos textos. Por fim, na Seção 4.4, são explicados os métodos utilizados para implementar o aprendizado de máquina em abordagens supervisionadas e semi-supervisionadas nos experimentos.

A Figura 4.1 ilustra o fluxograma contendo as etapas realizadas no decorrer dos experimentos para cada conjunto de *tweets*. Primeiramente, foi realizada a coleta dos *tweets*, seguida por sua anotação manual (da Base Português apenas, uma vez que a Base Inglês já estava rotulada). Então, foram realizados o pré-processamento e a caracterização dos textos para que fosse possível aplicar os modelos de classificação supervisionados e semi-supervisionados. Por fim, os resultados são analisados a partir das métricas de avaliação utilizadas.

4.1 Coleta de dados

Foram utilizadas duas bases de dados diferentes nos experimentos deste trabalho. Uma em português construída e rotulada manualmente a partir de *tweets* relacionados à Universidade de Brasília, que será chamada de "Base Português". A segunda base de *tweets* é formada por 9684 *tweets* na língua inglesa utilizada no trabalho de Rosenthal et al. [31], chamada de "Base Inglês".

O Twitter disponibiliza uma biblioteca em Python chamada Tweepy[1] , pela qual é possível coletar dados do Twitter, desde usuários, localização ou seguidores, até os próprios *tweets*, filtrando por palavras-chave, datas, entre outros argumentos. Para usar esta

[1] `https://www.tweepy.org/`

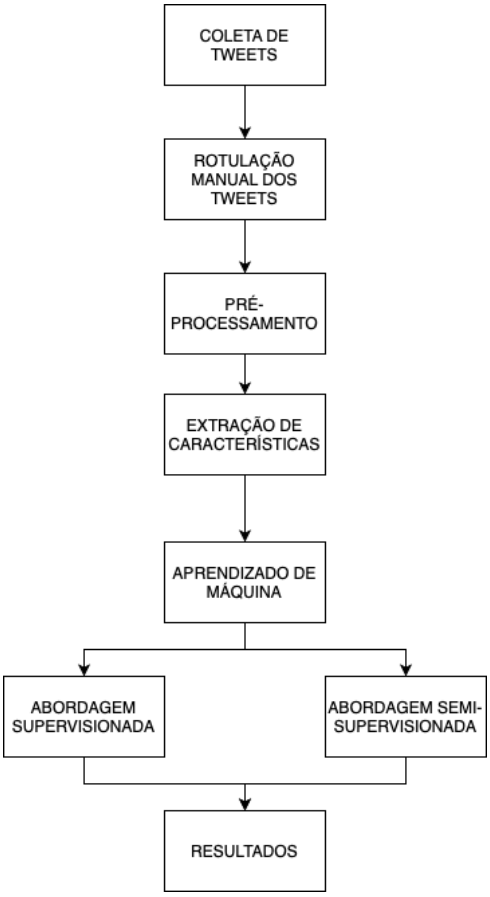

Figura 4.1: Fluxograma descrevendo as etapas da metodologia proposta.

biblioteca, é necessário realizar um cadastro de desenvolvedor no site da plataforma, explicando o motivo pelo qual os dados coletados serão usados. A requisição é analisada pela plataforma e então recebem-se *tokens* e chaves de acesso que são necessárias para autorizar o programa criado à acessar a API do Twitter. Utilizando esta ferramenta, primeiramente, foram pesquisados diversos perfis do Twitter relacionados à UnB, obtendo-se 23 perfis diferentes, entre eles, perfis oficiais da UnB (como *@unb_oficial*, ou *@dceunb*), perfis de centros acadêmicos (como *@cafarunb* ou *@cacomunb*), entre outros. Em seguida, foram obtidos os números identificadores, *ids*, dos seguidores de cada um destes perfis (exceto os que possuíam conta privada), chegando à lista de mais de 100 mil *ids*. Então, iniciou-se o processo de coleta de *tweets* de cada *id*, através da Tweepy. O foco do projeto é o desenvolvimento de um filtro com *tweets* que fossem relacionados à Universidade, sendo assim, foram selecionados apenas *tweets* que continham algumas palavras-chave previamente selecionadas, como “universidade”, “aula”, “trabalho”, "professor”, entre outras, assim como suas abreviaturas, como “prof”. O processo de coleta é demorado, uma vez que O Twitter limita a coleta de *tweets* levando em conta uma determinada quantidade de perfis visitados, sendo necessário interromper a coleta momentaneamente. Assim, um

Classificação dos Tweets		
Classificação	Número de *tweets* - Base Português	Número de *tweets* - Base Inglês
Positivo	475	3640
Negativo	683	1458
Neutro	804	4586
Total	1962	9684

Tabela 4.1: Tabela de número de rótulos classificados dos Tweets.

processo de busca dos 50 primeiros *tweets* de cerca de 1000 usuários levava cerca de quatro horas para ser realizado.

No total, foram coletados mais de 16.000 *tweets*, dos quais 2.000 aleatórios foram rotulados em positivo (1), negativo (2) ou neutro (0). Para diminuir o erro humano, cada *tweet* foi rotulado por três pessoas diferentes e o rótulo final era decidido com base no sentimento mais votado para cada *tweet*. Um total de 38 destes *tweets* foram rotulados como negativo, neutro e positivo por três diferentes pessoas. Por serem muito ambíguos, estes ficaram de fora dos experimentos que totalizou no final 1962 *tweets* na Base Português. Por sua vez, a Base Inglês já conta com 9684 *tweets* em inglês rotulados.

A Tabela 4.1 apresenta o número de rótulos positivos, negativos e neutros que as duas bases de *tweets* rotulados recebeu. A Tabela 4.2 apresenta uma amostra dos *tweets* e seus rótulos da Base Português, enquanto a Tabela 4.3 apresenta uma amostra dos *tweets* e seus rótulos da Base Inglês.

Amostra da Base Português rotulada	
Tweet	**Rótulo**
Alunos e professores da UFFS, Campus Cerro Largo, apresentam trabalho em Goiás	Neutro
Me irrita muito estudar coisa INÚTIL	Negativo
Que preguiça do cão de ir pra unb hj :/	Negativo
Venha estudar na PUC-Rio! Inscrições abertas para o vestibular!	Neutro
Eu to só enrolando pra fazer o meu trabalho	Negativo
felicidade do universitário é receber email do prof avisando que não vai ter aula das 08h	Positivo
poxa deu até vontade de estudar real	Positivo
INFEEERNO, passei duas horas fazendo a merda de um cálculo pra no final dá errado, d e s i s t o	Negativo
Melhor turma. Mozzato Atores e Modelos...	Positivo
Esses trabalhos infinitos, aiai :x	Negativo
Amanha é dia de prova. Vamos dormir que é mais um grande dia.	Positivo

Tabela 4.2: Amostra de Tweets rotulados da Base Português.

Amostra da Base Inglês rotulada	
Tweet	**Rótulo**
i want to watch The Grey... for the 16th time.	Positivo
FUCK Halloween. Fuck thanksgiving. Fuck my birthday. Fuck Christmasc the New Yearc Valentinesc St. Patricksc April fools. Fuck all of that.	Negativo
I really thought HISD didnt have school tomorrow because it was Christopher Columbus day.	Negativo
tomorrow gonna be hell yall. thoughc i am going to the golden modes in chapel hill!	Neutro
Project X is the best film ever made and someone with money needs to have onec like this Friday. #makeithappen	Positivo
Just landed in Las Vegas!!! So excited to start canvassing for Obama tomorrow with @schuy_g and @Kimbemo! :) #GetOutTheVote	Positivo
Poland criticize the 18th amendment. The new frame has been opened. None of the topic was on that earlier. #UPRLKA	Neutro
Cba with work tomorrow! #Boring #Blag	Negativo
Sticks and stones may break my bonesc but Rugby does it better!	Positivo
the only thing i may truly miss about lexington is shotos !!!	Negativo
Whos running the StrattonFaxon 20K in New Haven CT on Monday?	Neutro

Tabela 4.3: Amostra de Tweets rotulados da Base Inglês.

4.2 Pré-Processamento dos Dados

Na área de análise de sentimentos, quando os dados coletados são não-estruturados é importante que sejam tratados primeiramente, retirando informações redundantes e irrelevantes à classificação. Esta etapa do processo é conhecida como pré-processamento, ou limpeza de dados, e busca corrigir as inconsistências, preenchimento de informações, remoção de ruído e redundâncias, etc. [32].

Assim, para a Base Português, foram removidos *links* dos textos, *stop-words* da língua Portuguesa, como por exemplo, preposições e artigos, além de pontuações e caracteres não alfa-numéricos. Adicionalmente, foram retirados *tweets* que eram respostas, também conhecidos como *replies*, uma vez que muitas respostas não faziam sentido sem o contexto, dificultando o processo de rotulação. Todas as palavras de cada tweet foram convertidas em letras minúsculas, com a finalidade de padronizar o texto. A princípio, para a Base Português, foram usadas *stop-words* em português providas pela biblioteca NLTK[2], porém, posteriormente, foram feitos quatro testes com diversos conjuntos de *stop-words* encontrados online, e foi mantido o conjunto em que os classificadores obtiveram o melhor F1-Score. Duas técnicas de pré-processamento foram testadas na Base Português: lematização e stemização. Para testar a lematização, foi utilizada a biblioteca *spaCy*[3]. No entanto, não foi obtida nenhuma melhora na acurácia dos modelos de classificação treinados e testados, portanto os resultados não foram incluídos na experimentação. Já o teste da stemização provou-se bem sucedido para dois modelos, *Naive Bayes* e *Label Propagation*. O algoritmo escolhido foi o *RSLP Stemmer*[4] da biblioteca NLTK.

O pré-processamento de dados da Base Inglês foi semelhante em relação à Base Português, em que foram retiradas do texto *stop-words*, *links*, pontuações e caracteres especiais. As técnicas de stemização e lematização também foram consideradas para esta etapa do processo. Para isso, foram testados dois *stemmers* diferentes para o conjunto de dados: *Lancaster Stemmer*[5] e o *Porter Stemmer*[6], ambos disponibilizados pela biblioteca NLTK. Já a técnica de lematização considerada também foi provida pela biblioteca NLTK, a *WordNetLemmatizer*[7]. Para cada modelo de classificação, foram testadas as técnicas de stemização e lematização em conjunto e separadamente nos conjuntos de dados do experimento. Foi observado que os modelos de classificação apresentavam maiores F-Scores apenas em algumas ocasiões em que a técnica de lematização era aplicada. Sendo assim,

[2] `http://www.nltk.org/`
[3] `https://spacy.io/models/pt`
[4] `https://www.nltk.org/_modules/nltk/stem/rslp.html`
[5] `https://www.nltk.org/_modules/nltk/stem/lancaster.html`
[6] `https://www.nltk.org/_modules/nltk/stem/porter.html`
[7] `https://www.nltk.org/_modules/nltk/stem/wordnet.html`

na etapa de pré processamento, somente a técnica de lematização foi utilizada, também nos modelos de classificação *Naive Bayes* e *Label Propagation*.

4.3 Extração de Características

Para viabilizar o emprego dos modelos de classificação, é necessário gerar uma representação numérica para os textos. Para este fim, foram consideradas as técnicas *TF-IDF* e *Bag of Words* da biblioteca *Scikit Learn* em Python, que realizam a transformação do texto em números de acordo com a frequência dos seus termos. Decidiu-se por utilizar a técnica *TF-IDF*, que realiza a transformação do texto em números de acordo com a frequência inversa dos seus termos, por ter sido mais comumente utilizada nos trabalhos relacionados estudados, como o estudo de Li et al. [27] e o trabalho de Hong et al. [28].

4.4 Aprendizado de Máquina

Nesta etapa, foram utilizados os modelos de classificação *SVM*, *Naive Bayes*, *KNN* e *Label Propagation* para treinar e classificar *tweets* das duas bases de dados coletadas. Primeiramente, a abordagem supervisionada foi implementada e testada em cada conjunto de dados para cada modelo de classificação. Em seguida, a abordagem semi-supervisionada foi implementada para chegar aos resultados. Neste projeto, foram abordadas duas categorias de aprendizado semi-supervisionado: o método *wrapper-based* [4], mais especificamente a representação de auto-treinamento, e o método baseado em grafos [4].

Decidiu-se pela utilização de quatro métodos diferentes, com o objetivo de comparar seus resultados: Para a abordagem de auto-treinamento, foram utilizados o SVM[8], o Naive Bayes[9] e o KNN[10]. O GridSearch[11] foi escolhido para encontrar os melhores parâmetros para análise no modelo de classificação SVM.

Existem diversas variações do modelo de classificação *Naive Bayes*. McCallum et al. [21] realizou experimentos para comparar as variações deste classificador, sendo concluído que a variação *Multinomial Naive Bayes Classifier* mostrou apresentar melhores resultados nos conjuntos de dados com características similares as das Bases Português e Inglês. Desta forma, esta variação foi escolhida para prosseguir com os experimentos deste trabalho. Para o modelo de classificação *KNN*, foram testados os números de vizinhos

[8] `https://scikit-learn.org/stable/modules/svm.html`

[9] `https://scikit-learn.org/stable/modules/naive_bayes.html`

[10] `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html`

[11] `https://scikit-learn.org/stable/modules/grid_search.html`

entre 2 e 9 e foi decidido trabalhar com o número de vizinhos $K = 6$, por apresentar avaliações com maiores taxas de F-Score nos classificadores com esta configuração.

Para a abordagem baseada em grafos, usou-se o método *Label Propagation*[12] de aprendizado semi-supervisionado. A biblioteca *scikit-learn* fornece dois modelos de *Label Propagation*, o *LabelPropagation*[13] e o *LabelSpreading*[14]. Ambos possuem seu funcionamento baseado na construção de um grafo de similaridade sobre todos os itens do conjunto de dados de entrada, e se diferenciam nas modificações da matriz de similaridade e no efeito de fixação (*clamping effect*) nas distribuições de rótulos. Ambos modelos foram testados nos experimentos e decidiu-se por utilizar o modelo *LabelSpreading* com os parâmetros $kernel =$ "*knn*" e $alpha = 0.5$, pois, nas avaliações de acurácia dos classificadores, estas configurações apresentaram maiores taxas de F1-Score. A utilização do parâmetro $kernel =$ "*knn*" produz uma matriz esparsa muito mais amigável à memória, que pode reduzir drasticamente os tempos de execução, e o parâmetro *alpha* é o fator de fixação (*clamping factor*) que especifica a quantidade relativa que uma instância deve adotar as informações de seus vizinhos em oposição ao seu rótulo inicial, variando de zero a 1. O parâmetro $alpha = 0$ significa manter as informações iniciais do rótulo, enquanto $alpha = 1$ significa substituir todas as informações iniciais.

4.4.1 Auto-Treinamento

A obtenção de classificadores na categoria de modelos de auto-treinamento foi dividida em três etapas, cada uma gerando um classificador, como pode-se ver na Figura 4.2. Na primeira, foram separados 36% *tweets* para teste, chamados de "conjunto TESTE" e 64% para treinamento chamados de "conjunto TREINAMENTO", em cada uma das bases de *tweets*, gerando o primeiro classificador supervisionado. Em seguida, parte dos rótulos do conjunto TREINAMENTO foi ocultada, para que o restante dos *tweets* com rótulos mantidos fossem utilizados para treinar um classificador auxiliar que possui o objetivo de estimar o rótulo das demais instâncias (cujos rótulos foram ocultados). Para fins de experimentação, foi aplicada uma técnica utilizada no trabalho de Silva et al. [4], com objetivo de analisar como os classificadores iriam de comportar de acordo com a quantidade de rótulos usados em seu treino na abordagem semi-supervisionada, variando a proporção de *tweets* com rótulos ocultos a cada 10% em cada experimento, chegando enfim a 90% de rótulos ocultos no último experimento. Por exemplo, ao ocultar 10% dos rótulos do conjunto TREINAMENTO, os 90% *tweets* restantes foram utilizados para treinar e

[12] `https://scikit-learn.org/stable/modules/label_propagation.html`

[13] `https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelPropagation.html#sklearn.semi_supervised.LabelPropagation`

[14] `https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html#sklearn.semi_supervised.LabelSpreading`

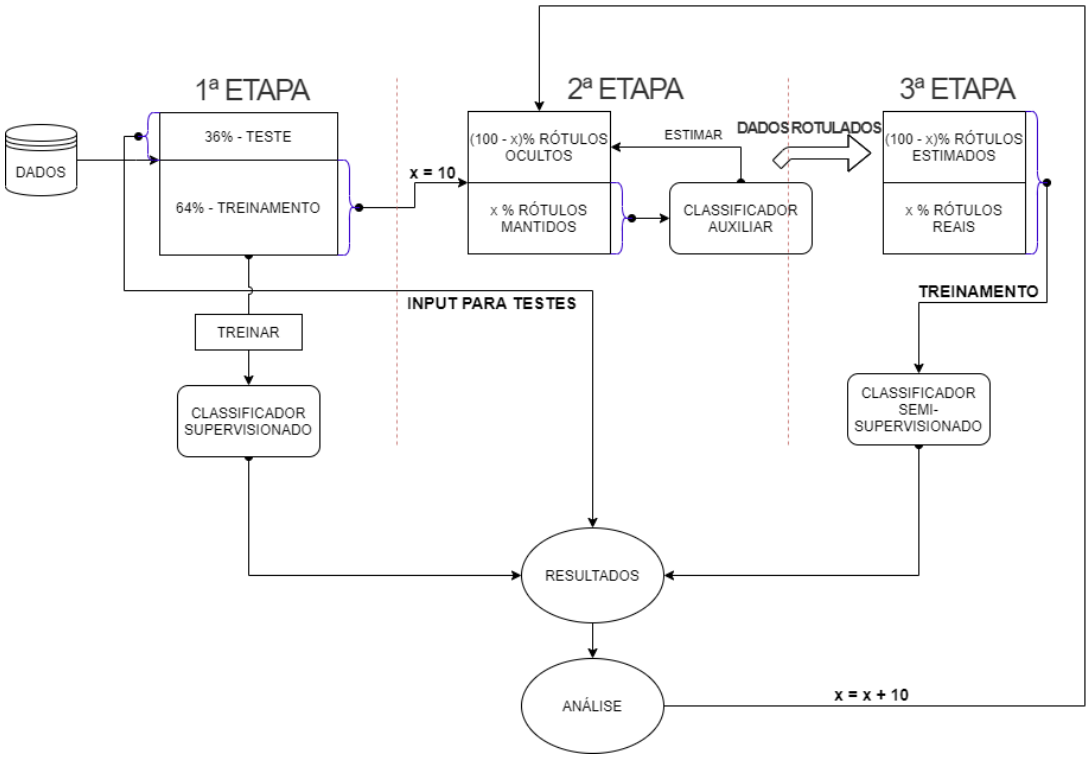

Figura 4.2: Fluxograma da Categoria de Auto-Treinamento.

gerar o classificador auxiliar que estimou os rótulos ocultos deste mesmo conjunto. Já na terceira etapa, foram utilizados os *tweets* de treinamento decorrentes da etapa anterior, com uma proporção de rótulos estimados e uma proporção de rótulos reais, para gerar o terceiro e último classificador, baseado em aprendizado semi-supervisionado.

Por fim, utilizando o conjunto TESTE como entrada, foram testados os classificadores supervisionado e semi-supervisionado, a fim de comparar os resultados gerados a partir das métricas de avaliação de performance dos modelos de classificação abordadas na Seção 2.3.5.

4.4.2 Baseado em Grafos

Já no método baseado em grafos, a metodologia era mais simples, uma vez que não era necessário criar um classificador auxiliar para estimar rótulos, como feito na segunda etapa do método anterior, resultando em apenas duas etapas conforme a Figura 4.3. Na primeira etapa, foram separados 36% *tweets* para teste, chamados de "conjunto TESTE" e 64% para treinamento, chamados de "conjunto TREINAMENTO", em cada uma das bases de *tweets*, gerando o primeiro classificador supervisionado a partir do algoritmo *Label Propagation*. Em seguida, seguindo a mesma técnica utilizada por Silva et al. [4], uma determinada porcentagem do conjunto TREINAMENTO foi oculta em cada execução,

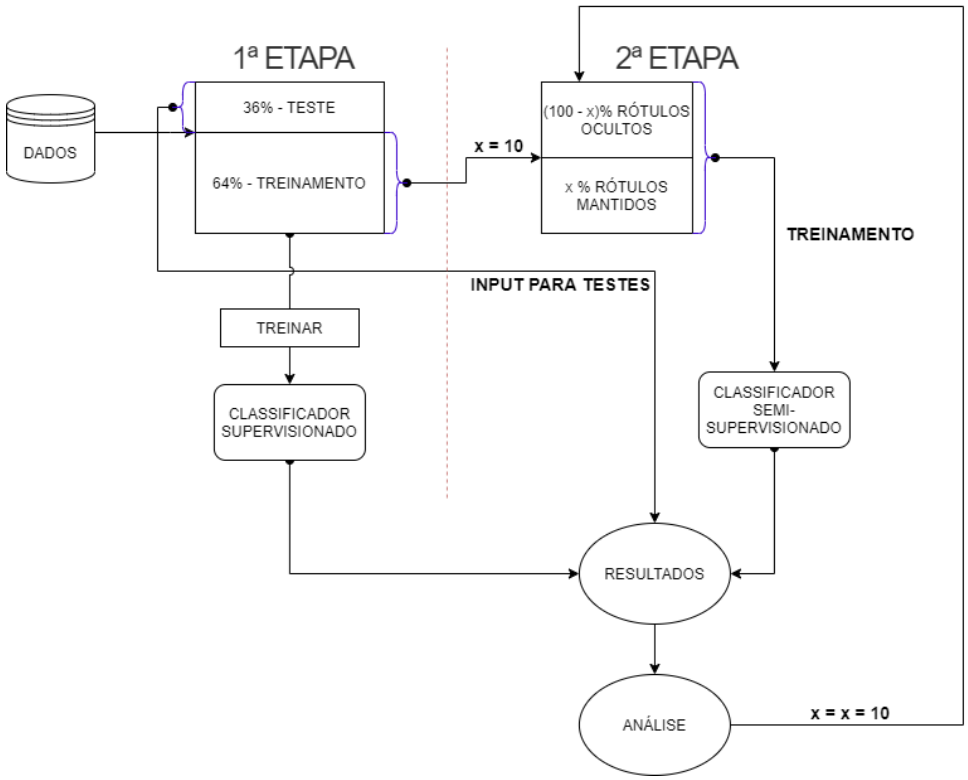

Figura 4.3: Fluxograma da Categoria Baseada em Grafos.

deixando apenas uma porcentagem da base de dados rotulada, de modo que o classificador fosse treinado com apenas 10% do conjunto TREINAMENTO rotulado, em seguida 20% do conjunto rotulado, e assim por diante, até ser executado o experimento com 90% do conjunto rotulado, treinando um novo classificador semi-supervisionado em cada execução.

Por fim, a partir do conjunto TESTE, foram testados os classificadores supervisionado e semi-supervisionado, com objetivo de comparar os resultados do método baseado em grafos a partir das métricas de avaliação de performance deste modelo de classificação abordados na Seção 2.3.5.

Capítulo 5

Resultados Experimentais

Este capítulo apresenta os resultados experimentais visando validar os métodos de classificação de sentimentos propostos. Na Seção 5.1, são discutidos os resultados provenientes dos experimentos realizados na Base Inglês, enquanto que na Seção 5.2, são abordados os resultados obtidos através dos experimentos realizados na Base Português.

Após cada execução dos experimentos descritos no Capítulo 4, foram gerados valores resultados provindos das métricas de avaliação de performance dos classificadores, ou seja, matrizes de confusão, valores de precisão, revocação e *F-Scores*. A partir destes valores, foi possível a geração de gráficos e tabelas para melhor visualização e análise destes resultados.

5.1 Base Inglês

Para a Base Inglês, os experimentos foram realizados em uma máquina com processador de 2.9 GHZ Dual-Core Intel Core i5, utilizando o Sistema Operacional MacOS Catalina versão 10.15.3. Os classificadores *Naive Bayes* levaram em média 2 segundos para treinar e prever o conjunto de *tweets* em cada configuração treino/teste do experimento, enquanto que os classificadores *KNN* (com o número de vizinhos $K = 6$, porque apresentaram melhores resultados nos testes realizados), levaram em média 12 minutos, *Label Propagation* levaram em média 30 minutos, e os classificadores *SVM* duraram em média 440 minutos. Cada experimento foi realizado dez vezes.

Os F-Scores resultantes dos experimentos realizados na Base Inglês podem ser vistos na Tabela 5.1. Já a Figura 5.1 mostra o gráfico dos F-Scores obtidos para cada um dos classificadores nos experimentos realizados na Base Inglês para cada porcentagem de *tweets* rotulados inicialmente. É possível concluir que o modelo de classificação *SVM* obteve os melhores resultados nos experimentos em que mais de 50% dos *tweets* eram rotulados

inicialmente, nos experimentos em que menos de 60% dos *tweets* eram inicialmente rotulados, o modelo de classificação semi-supervisionado baseado em grafos *Label Propagation* obteve maiores valores de F1-Score.

Apesar de possuir os maiores resultados em termos de acurácia, o modelo de classificação *SVM* foi o que também levou maior tempo para realizar o treinamento e predição dos dados, sendo significativamente maior que os demais, levando mais de 10 vezes o tempo que o segundo classificador mais demorado nestes experimentos, o *Label Propagation*.

Também é possível visualizar o F-Score obtido na abordagem supervisionada (em que 100% dos *tweets* são rotulados), e em todos os casos, esta abordagem apresentou as maiores taxas de F1-Score em relação às abordagens semi-supervisionadas. Os resultados superiores na abordagem supervisionada foram mais significativos nos modelos de classificação *Naive Bayes* e *SVM*. Os experimentos realizados nas abordagens semi-supervisionadas utilizando a Base Inglês não se provaram tão eficazes quanto as abordagens supervisionadas nos resultados, porém se mostraram uma abordagem apropriada para tarefas de classificação em que poucos dados estão rotulados. No entanto, de maneira geral neste experimento, quanto maior o número de dados inicialmente rotulados, maiores foram os valores de F1-Score obtidos pelos classificadores.

A partir das matrizes de confusão de cada experimento, foi possível criar os gráficos apresentados nas Figuras 5.2 - 5.11, que contêm a porcentagem de acertos de cada possível rótulo dos *tweets* (Positivo, Negativo ou Neutro) para cada porcentagem de *tweets* inicialmente rotulados para a Base Inglês. Fica evidente que todos os modelos de classificação obtiveram melhores taxas de acertos para classificar *tweets* neutros, comparados aos demais neste conjunto de dados, logo conclui-se que os classificadores treinados aprenderam melhor os padrões de *tweets* neutros do que as outras polaridades. Um possível indício que pode explicar este fenômeno é o fato de existir uma maior quantidade de *tweets* rotulados como neutro no conjunto de dados, obtendo-se maior quantidade de exemplos de treinamento para estimar este rótulo específico, ou seja, a representatividade da classe neutra no conjunto de *tweets* é um fator importante que pode levar à tais resultados. Outro indício a ser apontado é que os *tweets* neutros não possuem uma opinião forte (tanto positivas quanto negativas) nos seus textos, e a representação TF-IDF não captura o contexto. Ademais, outro fato preponderante é que os *tweets* possuem assuntos diversos, não sendo relacionados com contextos específicos.

Na competição SemEval2016, 34 participantes utilizaram esta base para classificação de *tweets* em positivo, negativo ou neutro (*subtask A: Message polarity classification*). Na avaliação dos classificadores, foram obtidos F-Scores variando entre 0.303 e 0.633. Dentre as abordagens utilizadas pelos participantes, 5 dos 10 sistemas com maior valor de F-Score utilizaram redes neurais de aprendizado profundo [33].

F-Score dos Classificadores - Base Inglês				
% de *Tweets* Rotulados	SVM	*Naive Bayes*	KNN	*Label Propagation*
10%	0.415	0.415	0.395	0.442
20%	0.416	0.415	0.415	0.461
30%	0.424	0.415	0.412	0.475
40%	0.434	0.415	0.424	0.483
50%	0.450	0.415	0.438	0.492
60%	0.516	0.415	0.458	0.489
70%	0.548	0.420	0.478	0.509
80%	0.572	0.436	0.486	0.512
90%	0.588	0.465	0.493	0.512
100% (Abordagem Supervisionada)	0.625	0.545	0.511	0.528

Tabela 5.1: F-Scores dos classificadores treinados por meio de aprendizado semi-supervisionado e supervisionado para a Base Inglês.

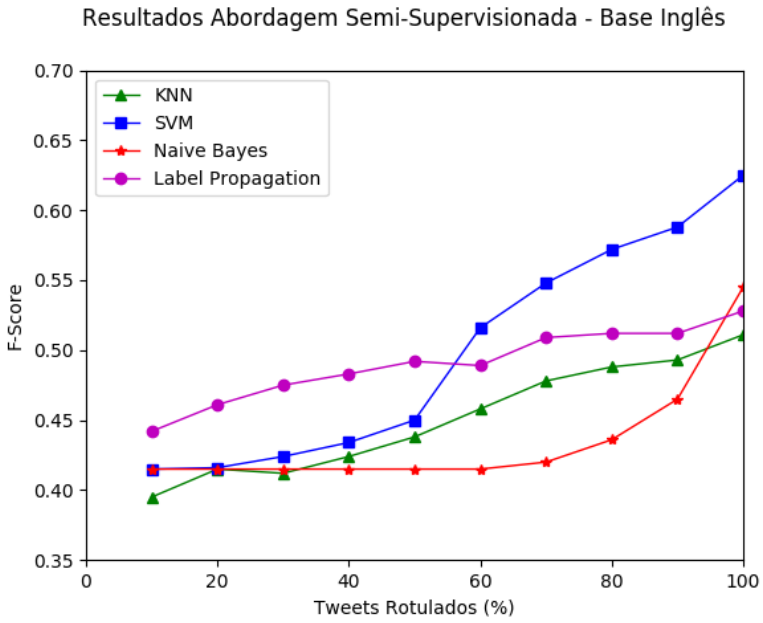

Figura 5.1: F-Score dos modelos de classificação para abordagens semi-Supervisionada e supervisionada.

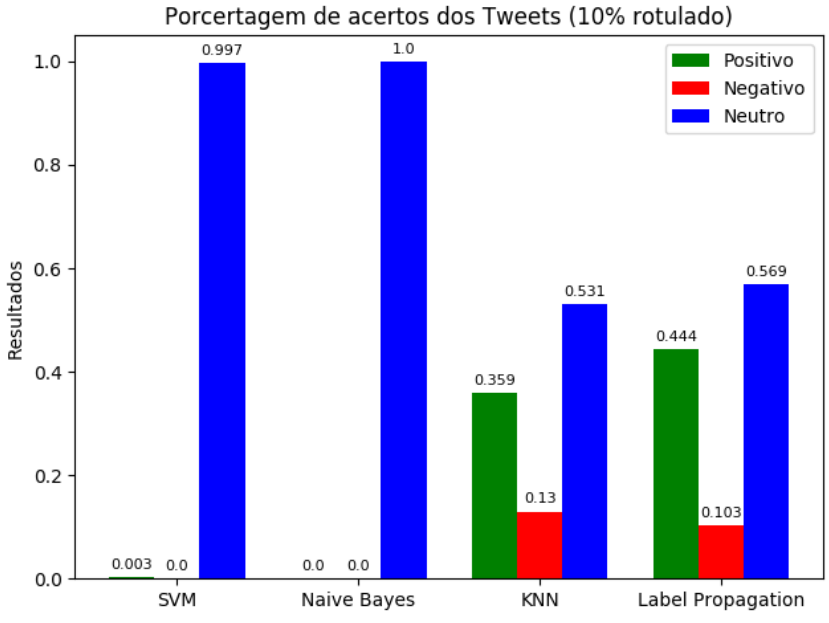

Figura 5.2: Porcentagem de acerto de *tweets* em conjunto com 10% de *tweets* rotulados para Base Inglês.

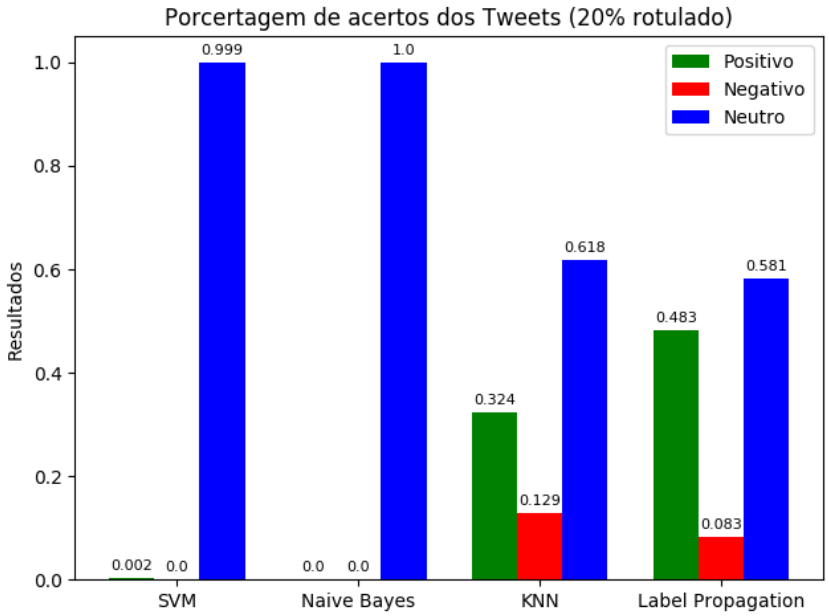

Figura 5.3: Porcentagem de acerto de *tweets* em conjunto com 20% de *tweets* rotulados para Base Inglês.

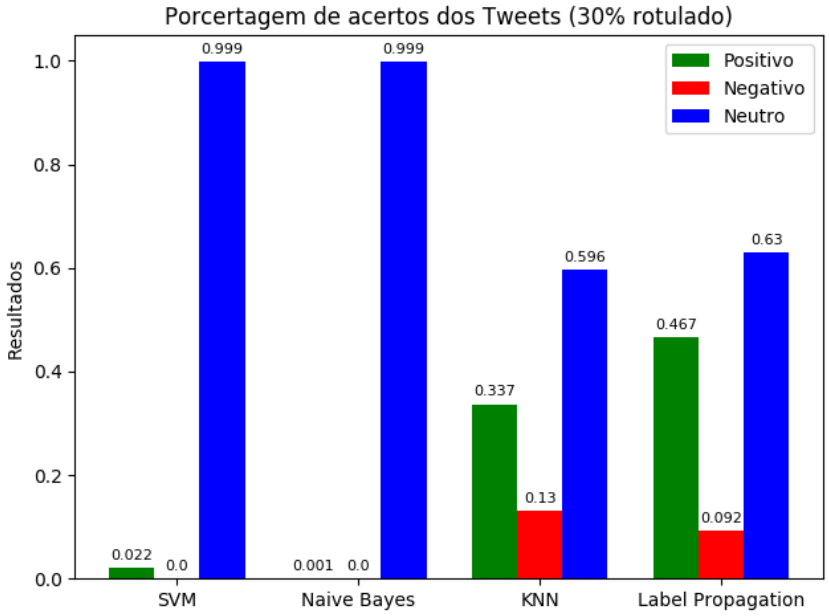

Figura 5.4: Porcentagem de acerto de *tweets* em conjunto com 30% de *tweets* rotulados para Base Inglês.

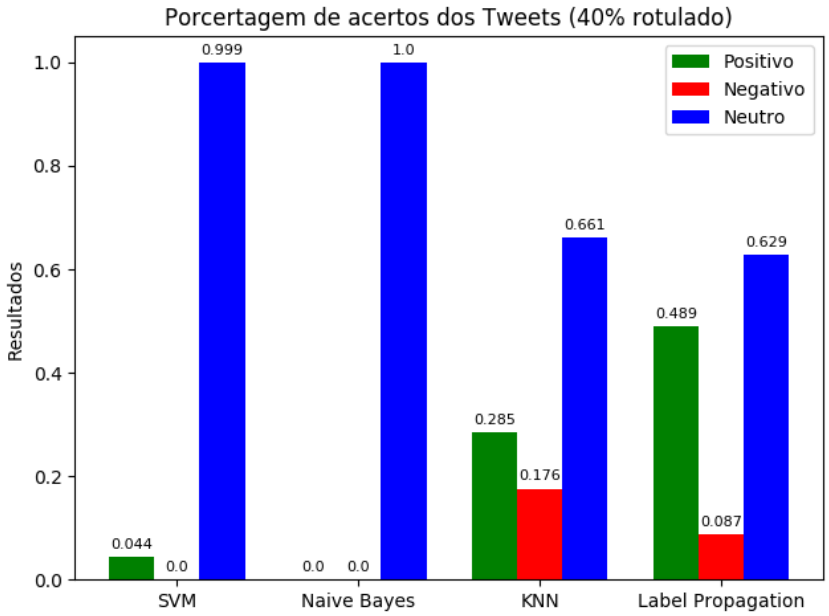

Figura 5.5: Porcentagem de acerto de *tweets* em conjunto com 40% de *tweets* rotulados para Base Inglês.

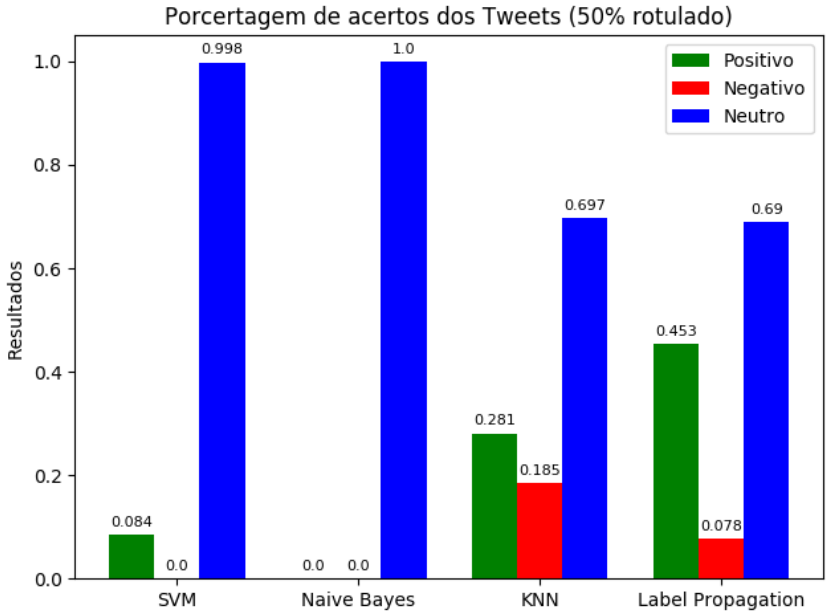

Figura 5.6: Porcentagem de acerto de *tweets* em conjunto com 50% de *tweets* rotulados para Base Inglês.

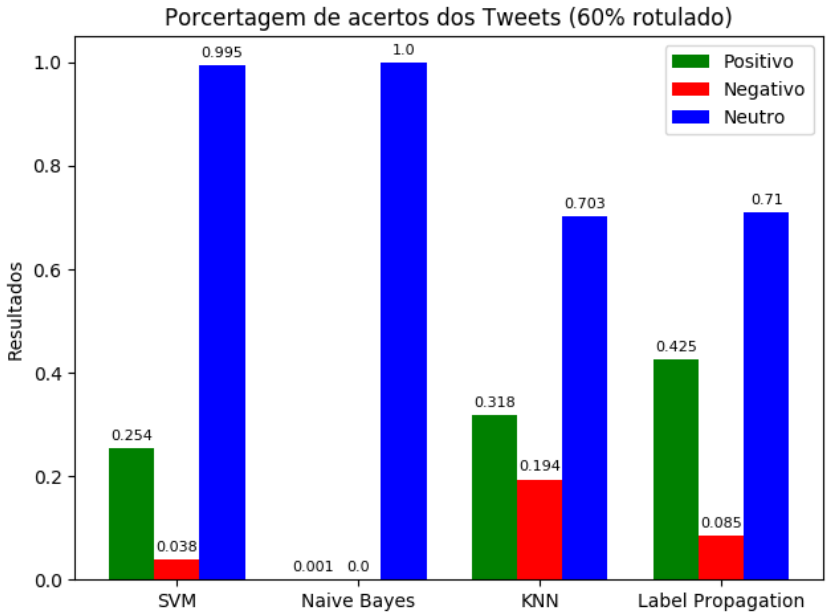

Figura 5.7: Porcentagem de acerto de *tweets* em conjunto com 60% de *tweets* rotulados para Base Inglês.

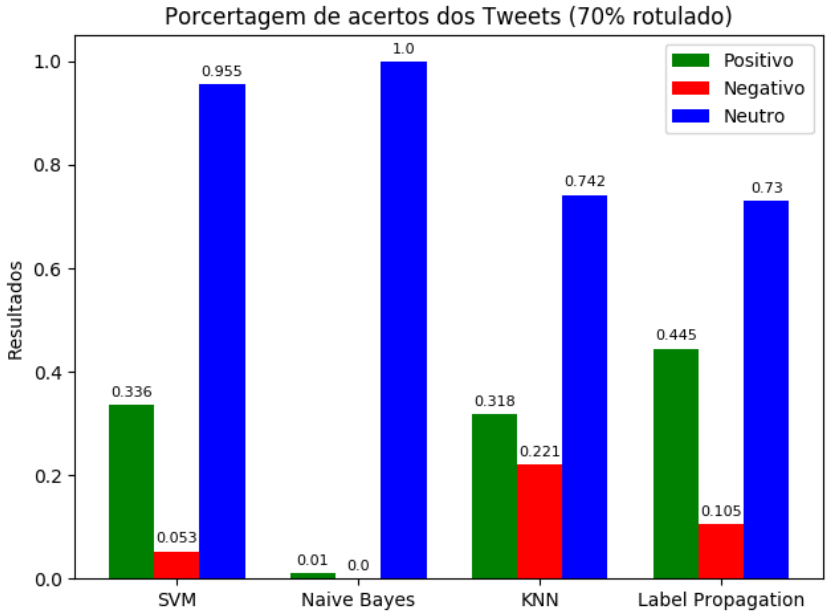

Figura 5.8: Porcentagem de acerto de *tweets* em conjunto com 70% de *tweets* rotulados para Base Inglês.

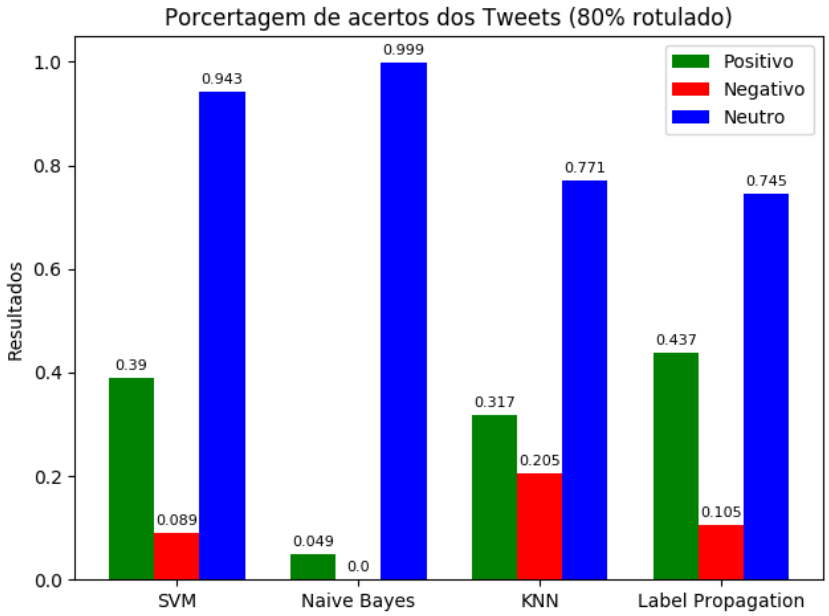

Figura 5.9: Porcentagem de acerto de *tweets* em conjunto com 80% de *tweets* rotulados para Base Inglês.

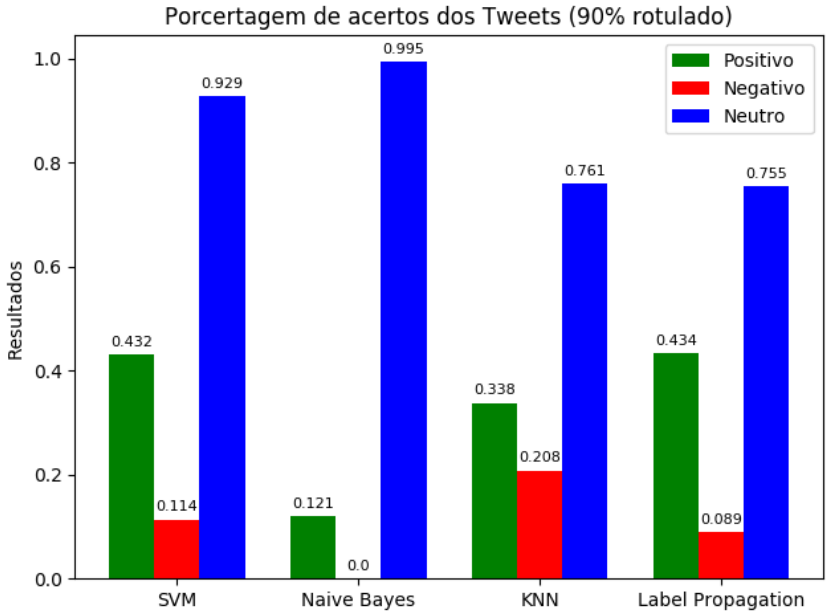

Figura 5.10: Porcentagem de acerto de *tweets* em conjunto com 90% de *tweets* rotulados para Base Inglês.

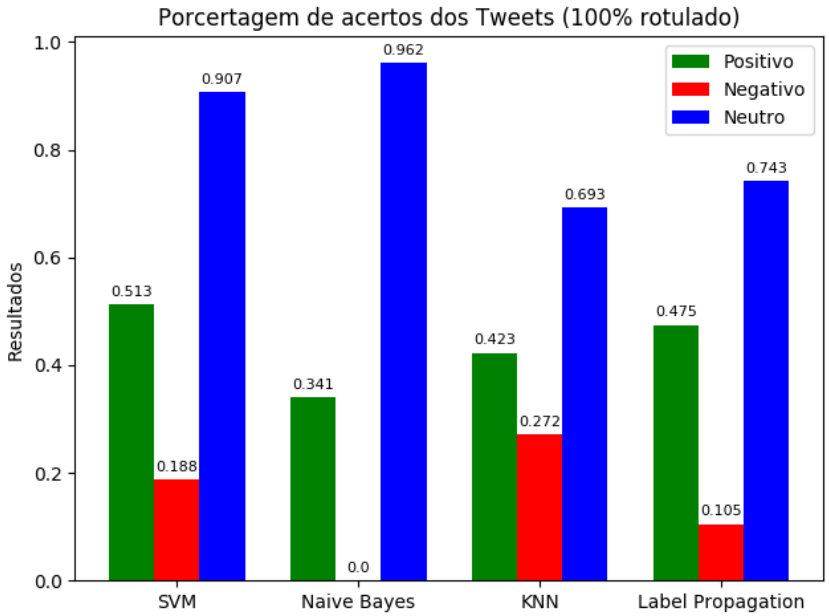

Figura 5.11: Porcentagem de acerto de *tweets* em conjunto com 100% de *tweets* rotulados para Base Inglês.

5.2 Base Português

Para a Base Português, a máquina utilizada para processar os experimentos possui as seguintes configurações: Intel Core i7-5500U 2.4GhZx4, utilizando o Sistema Operacional Ubuntu versão 16.04 LTS. Semelhante à Base Inglês, cada experimento foi executado pelo menos dez vezes, os *F-Scores* foram calculados e os resultados de acerto de cada categoria, *Positivo, Negativo e Neutro*, foram armazenados em matrizes de confusão.

O tempo para a realização do treino e da previsão de cada modelo foram medidos em todos os experimentos. O algoritmo que concluiu o treino e a previsão mais rápido foi o *Naive Bayes* com uma média abaixo de 2 segundos para cada configuração treino/teste do experimento. O KNN (com o número de vizinhos $K = 6$, por ter apresentado melhores resultados nos testes realizados) exibiu tempos com uma média abaixo dos 30 segundos. O modelo de propagação de rótulos demorou em média 1 minuto para concluir cada experimento. O SVM teve uma média de 10 minutos para concluir cada experimento, sendo o modelo mais lento em todos os testes realizados. É importante observar que a Base Português possui menor quantidade de *tweets* que a Base Inglês e o tamanho do conjunto de dados é muito relevante para se comparar tempo de processamento.

A Figura 5.12 mostra uma comparação de todos os modelos de classificação usados no experimento, mostrando seu desempenho de acordo com a porcentagem de *tweets* rotulados variando de 10% a 90%. O gráfico mostra claramente a superioridade nos valores de F1-Score do classificador SVM para todos os casos a partir de 20% dos *tweets*

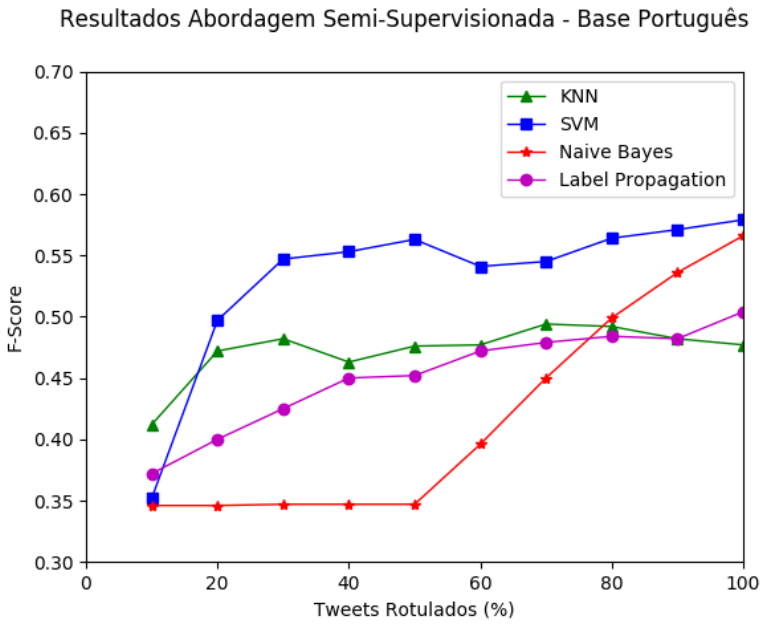

Figura 5.12: F-Score dos modelos de classificação para abordagens semi-Supervisionada e supervisionada para a Base Português.

rotulados. O KNN é o algoritmo que se mantém mais estável durante os experimentos com valores de F-Score entre 40% e 50%. O algoritmo *Label Propagation* apresenta o resultado mais consistente com o esperado antes da realização dos experimentos, um crescimento na acurácia de acordo com o aumento dos *tweets* rotulados. Os testes realizados com o modelo *Naive Bayes* se mantiveram idênticos entre as porcentagens de 10% e 50%, obtendo o pior F1-Score de todo o experimento, porém a partir dos 60% de *tweets* rotulados, o F1-Score cresce consideravelmente mostrando que esse algoritmo é muito eficiente para conjuntos de *tweets* que possuam uma parte significativa dos dados rotulados.

O resultado do F-Score de cada modelo conforme a quantidade de *tweets* rotulados é apresentado na Tabela 5.2. Os resultados apresentados na Tabela 5.2 foram os mesmos empregados no gráfico de comparação dos modelos na Figura 5.12.

Foram elaborados dez gráficos baseados nos valores das matrizes de confusão de cada modelo e a porcentagem dos *tweets* rotulados. As matrizes de confusão apresentaram dados relevantes sobre os experimentos não apenas entre os modelos, mas também numa comparação com a Base Inglês. Conforme mencionado anteriormente, a maior parte dos *tweets* da Base Inglês foi classificada como neutra, independentemente do modelo de classificação. As Figuras 5.13 - 5.22 mostram que a Base Português apresenta a maior parte dos *tweets* sendo classificados como neutros, porém os classificadores acertaram mais os *tweets* negativos.

Uma vez que haviam mais *tweets* classificados como neutro na base Português do que negativos ou positivos, podemos concluir que não necessariamente uma maior quantidade

F-Score dos Classificadores - Base Português				
% de *Tweets* Rotulados	SVM	*Naive Bayes*	KNN	*Label Propagation*
10%	0.352	0.347	0.411	0.372
20%	0.497	0.347	0.472	0.401
30%	0.547	0.347	0.482	0.425
40%	0.553	0.347	0.463	0.450
50%	0.563	0.347	0.476	0.452
60%	0.568	0.396	0.477	0.472
70%	0.563	0.450	0.494	0.479
80%	0.564	0.499	0.493	0.484
90%	0.571	0.536	0.483	0.482
100% (Abordagem Supervisionada)	0.580	0.567	0.477	0.504

Tabela 5.2: Tabela de F-Scores dos Classificadores Semi-Supervisionados e Supervisionado para Base Português.

de *tweets* classificadas em um rótulo específico na base significa maiores taxas de acertos deste rótulo na classificação do modelo. Um ponto importante a ser mencionado é que na Base Português, haviam 40% mais *tweets* rotulados como negativo em relação aos *tweets* rotulados como positivo, indicando que a opinião dos usuários acerca de assuntos relacionados a universidade tende a ser mais negativa do que positiva. Este fato também pode ter influenciado nos resultados das taxas de acerto dos modelos de classificação para a Base Português.

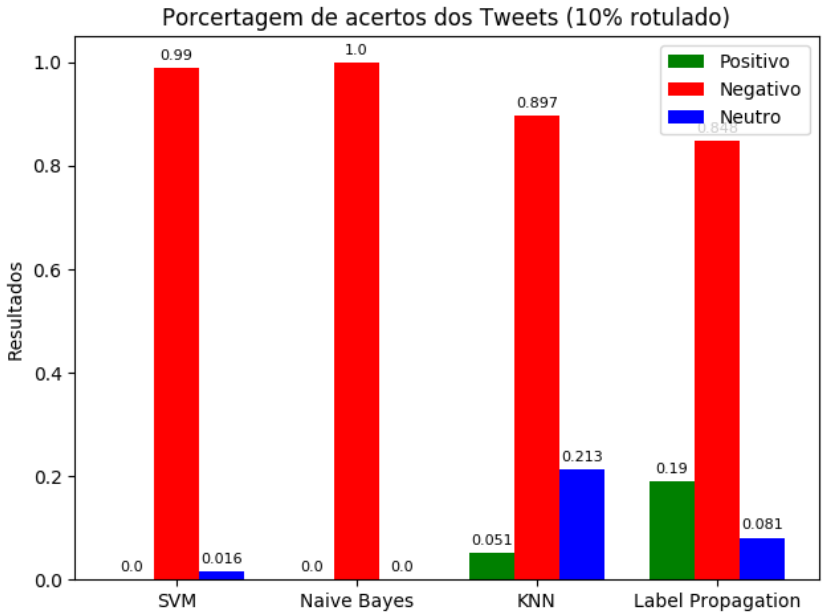

Figura 5.13: Porcentagem de acerto de *tweets* em conjunto com 10% de *tweets* rotulados para Base Português.

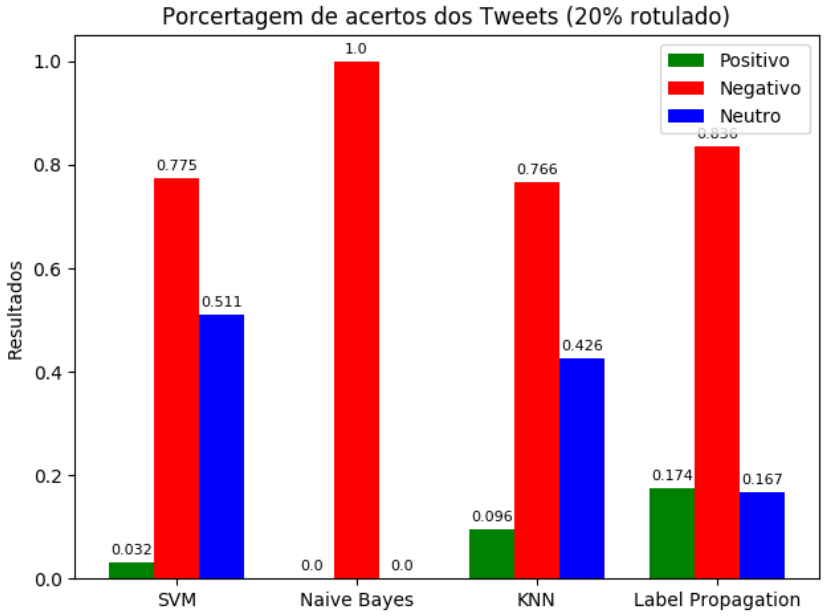

Figura 5.14: Porcentagem de acerto de *tweets* em conjunto com 20% de *tweets* rotulados para Base Português.

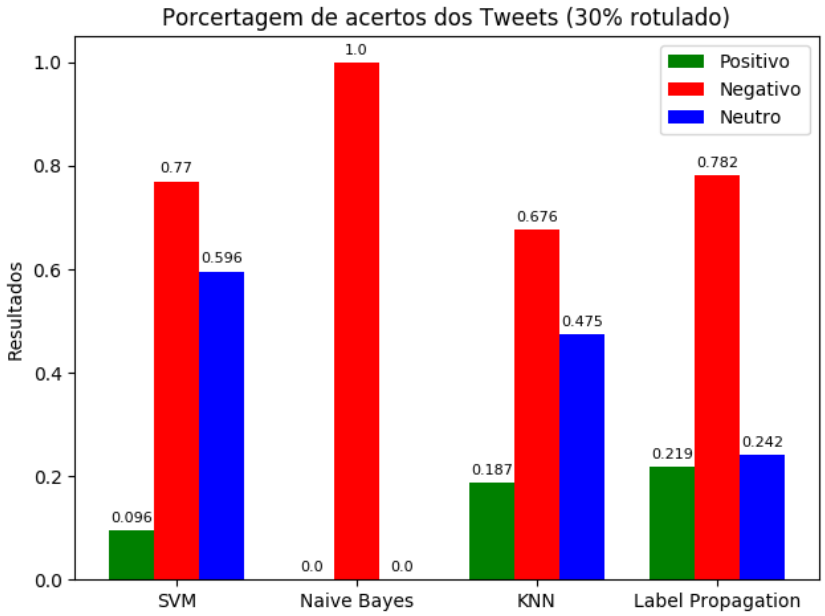

Figura 5.15: Porcentagem de acerto de *tweets* em conjunto com 30% de *tweets* rotulados para Base Português.

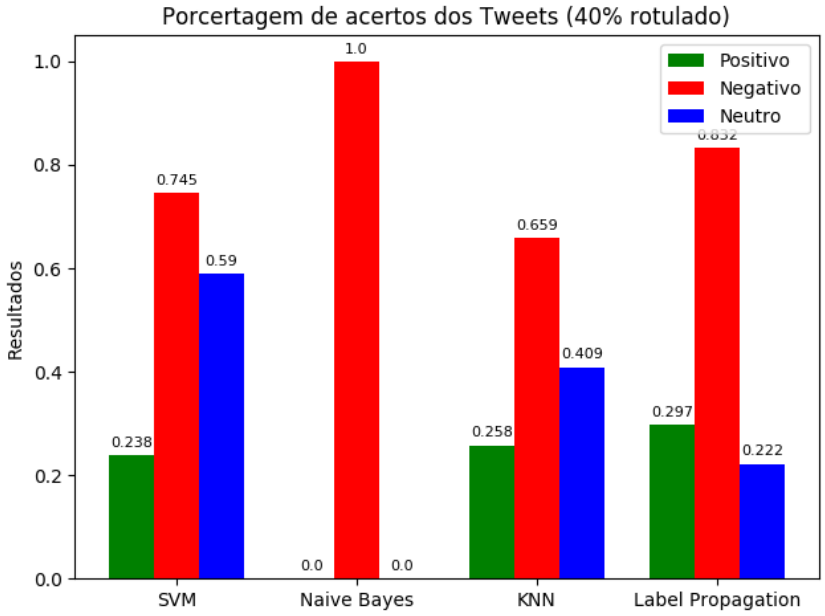

Figura 5.16: Porcentagem de acerto de *tweets* em conjunto com 40% de *tweets* rotulados para Base Português.

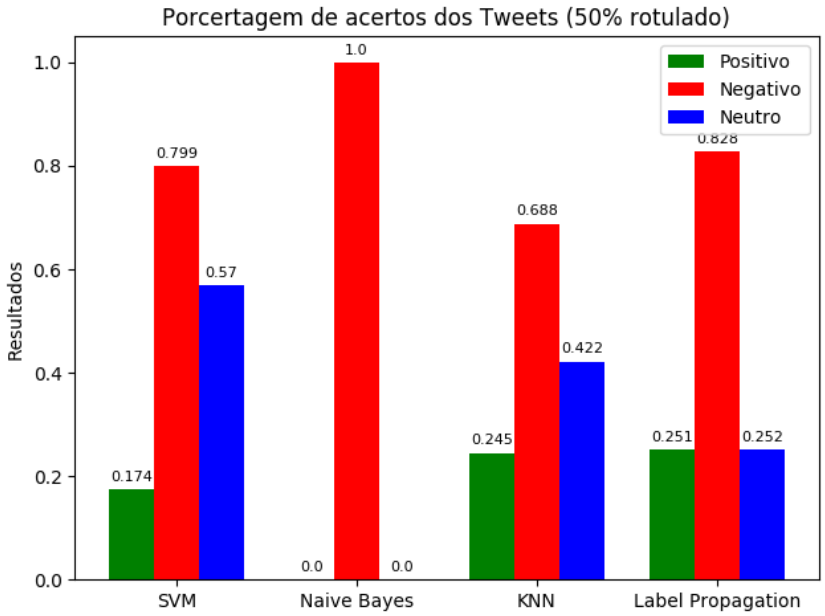

Figura 5.17: Porcentagem de acerto de *tweets* em conjunto com 50% de *tweets* rotulados para Base Português.

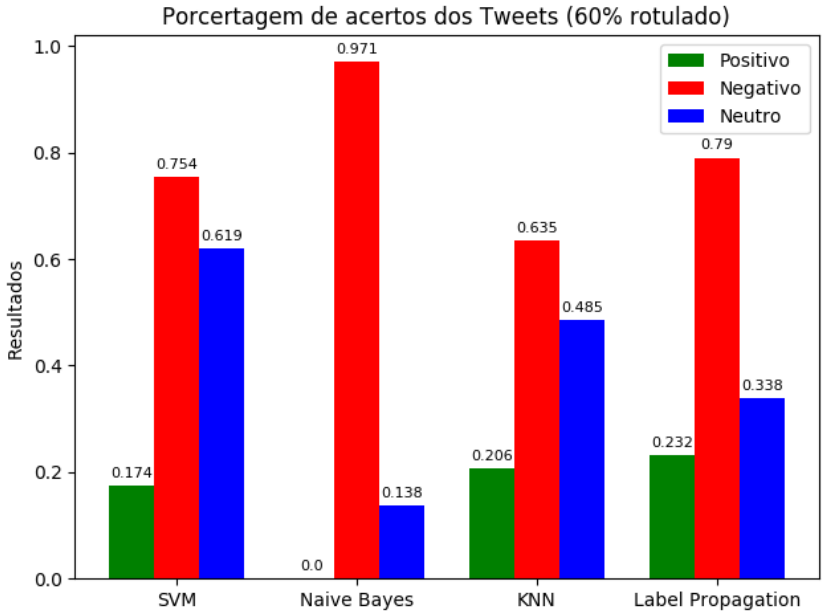

Figura 5.18: Porcentagem de acerto de *tweets* em conjunto com 60% de *tweets* rotulados para Base Português.

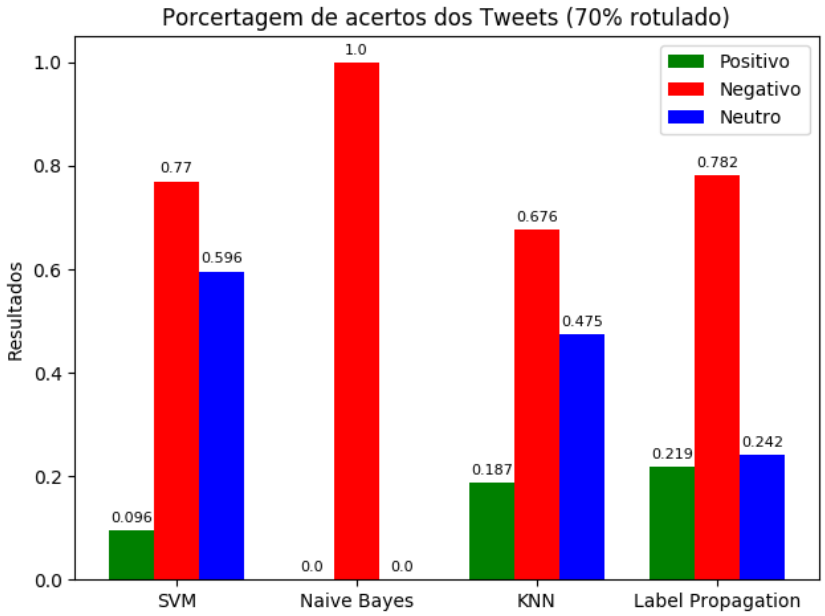

Figura 5.19: Porcentagem de acerto de *tweets* em conjunto com 70% de *tweets* rotulados para Base Português.

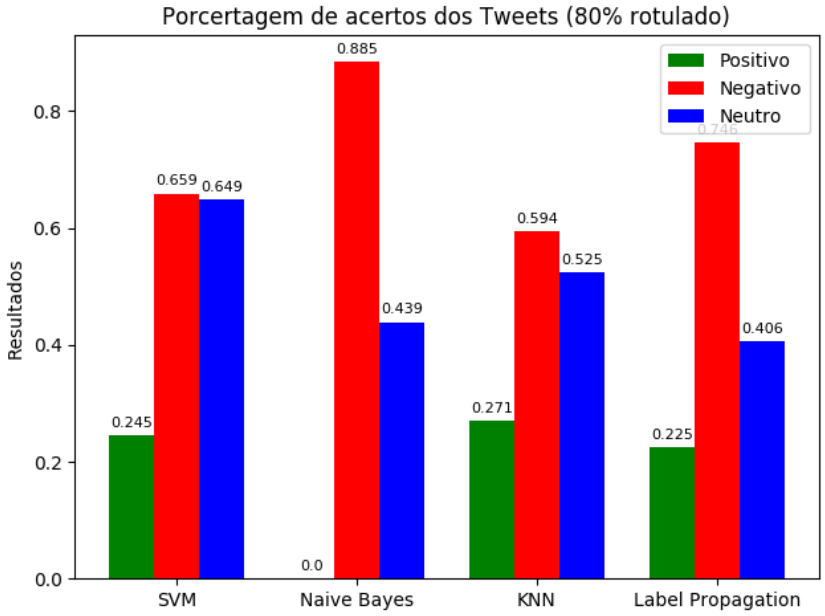

Figura 5.20: Porcentagem de acerto de *tweets* em conjunto com 80% de *tweets* rotulados para Base Português.

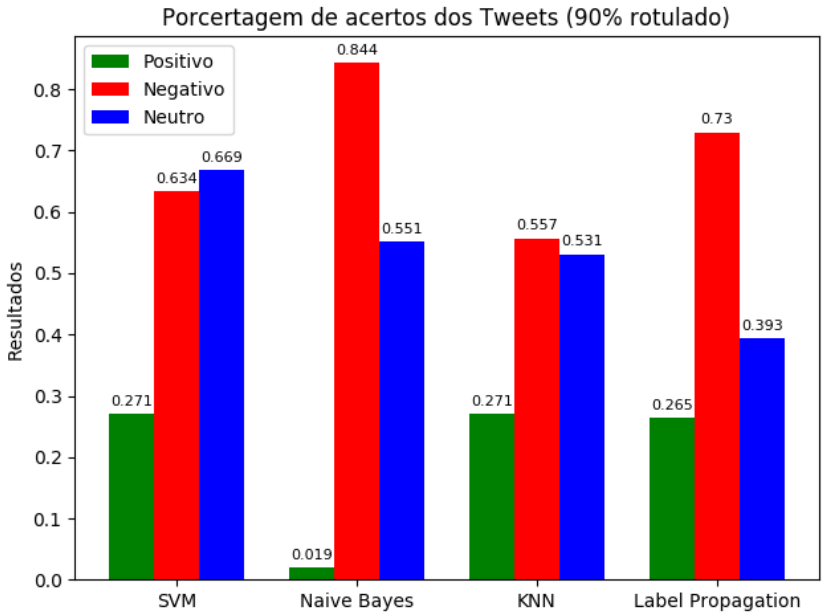

Figura 5.21: Porcentagem de acerto de *tweets* em conjunto com 90% de *tweets* rotulados para Base Português.

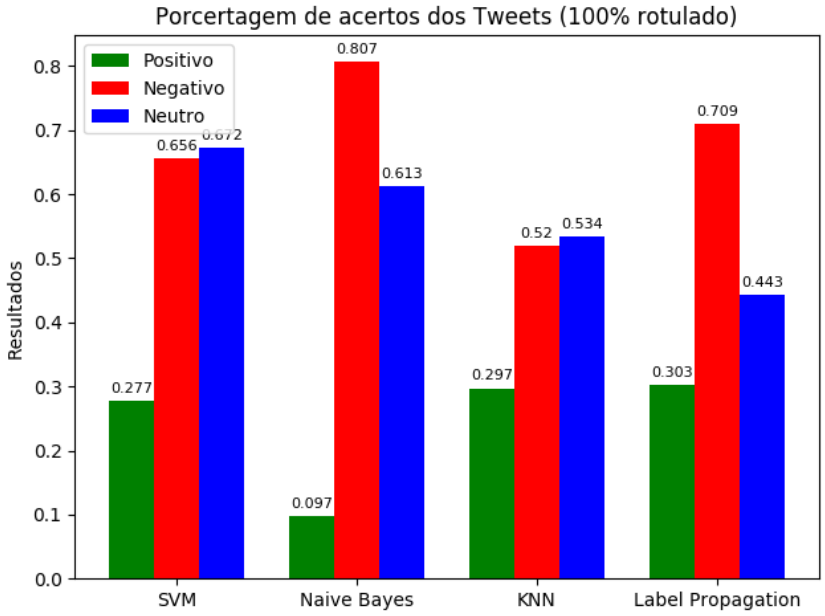

Figura 5.22: Porcentagem de acerto de *tweets* em conjunto com 100% de *tweets* rotulados para Base Português.

Capítulo 6

Conclusão

6.1 Considerações Finais

Este trabalho apresentou uma comparação de diferentes métodos de aprendizado de máquina semi-supervisionados em relação à métodos supervisionados, aplicados à análise de sentimentos para detecção e classificação de polaridades. Os experimentos foram realizados em dois conjuntos de dados de *tweets*, um na língua Portuguesa, chamado de "Base Português" e outro na língua inglesa, chamado de "Base Inglês". A Base Inglês, provida pelo trabalho de Rosenthal et al. [31], dispunha de 9684 *tweets* em inglês rotulados em positivo, negativo ou neutro, enquanto que a Base Português foi criada especificamente para este trabalho, e dispõe de 1962 *tweets* em português relacionados à Universidade de Brasília rotulados também em positivo, negativo e neutro. A Base Português construída e rotulada pode ser considerada a primeira contribuição deste trabalho.

Primeiramente, o tema de análise de sentimentos foi contextualizado, e os problemas sobre a falta de estudos realizados em conjuntos de dados na língua Portuguesa e a dificuldade de se rotular conjuntos de dados no geral foram expostos. Em seguida, foram explorados conceitos relacionados à análise de sentimento, que envolveriam a metodologia presente nesta monografia, textos, mineração de textos e aprendizado de máquina. Subsequentemente, quatro trabalhos relacionados à analise de sentimentos foram abordados, podendo-se levantar informações referente às técnicas mais utilizadas para aprendizado de máquina nesta área e reforçando a dificuldade de se encontrar trabalhos com experimentos na língua Portuguesa.

Decidiu-se, então, a metodologia proposta a ser imposta nos experimentos deste trabalho. Os meios utilizados para coleta e rotulação de dados foram explicados, da mesma forma que o processo de pré-processamento e caracterização de tais dados também foi discutida. Em seguida, foram expostos os modelos de classificação a serem utilizados nos experimentos, e como o aprendizado de máquina semi-supervisionado seria abordado.

Por fim, foi-se possível chegar aos resultados através da análise dos dados gerados pelas métricas de avaliação de performance dos classificadores.

Concluiu-se que, nos experimentos realizados na Base Inglês, o modelo de classificação SVM apresentou os melhores resultados nas situações em que pelo menos 60% do conjunto de dados utilizado era inicialmente rotulado, enquanto que em situações onde mais da metade do conjunto de dados não era rotulado, o modelo de classificação semi-supervisionado baseado em grafos *Label Propagation* retornou melhores resultados. Também foi possível concluir, através da análise dos dados provindos das matrizes de confusão resultantes dos experimentos, que *tweets* rotulados como neutros possuíam uma taxa de acerto significativamente maior que os demais.

Já nos experimentos realizados na Base Português, foi concluído que o modelo de classificação SVM se manteve com as maiores taxas de F-Score em todos os experimentos, com exceção na ocasião onde apenas 10% do conjunto de dados utilizado era rotulado inicialmente, na qual o modelo KNN se sobressaiu. Nota-se que o modelo de classificação KNN apresentou resultados melhores na abordagem semi-supervisionada com mais de 70% de *tweets* inicialmente rotulados, do que na abordagem supervisionada onde 100% dos *tweets* eram rotulados. Apesar de não possuir F-Scores tão altos quando o classificador SVM, o modelo de classificação *Naive Bayes* obteve resultados próximos nos experimentos com mais de 80% de dados inicialmente rotulados, sendo uma opção interessante por se tratar do modelo de classificação com melhor desempenho em termos de tempo de processamento para treinar e realizar a classificação de *tweets* de teste. Também foi possível concluir, através da análise dos dados provindos das matrizes de confusão resultantes dos experimentos, que *tweets* rotulados como negativos possuíam uma taxa de acerto significativamente maior que os demais nos experimentos realizados na Base Português.

Por fim, pode-se concluir na avaliação dos modelos de classificação que o SVM obteve maiores taxas de F-Score nos experimentos com a Base Português e com a Base Inglês, que continha maior quantidade de *tweets*. Para conjuntos de dados maiores, como no experimento da Base Inglês, o classificador *Label Propagation* se mostrou superior para abordagens semi-supervisionadas com pequenas taxas de dados rotulados inicialmente. Já em conjuntos de dados menores, como no experimento da Base Português, o classificador *Naive Bayes* apresentou resultados satisfatórios em abordagens supervisionadas, ou semi-supervisionadas com uma grande taxa de dados rotulados inicialmente.

6.2 Trabalhos Futuros

É possível aprimorar esta pesquisa em diversos âmbitos: para trabalhos futuros, seria interessante poder trabalhar com um conjunto de dados maior rotulado na língua Portuguesa para testar os experimentos. Sendo assim, continuar a rotulação dos *tweets* em português coletados seria uma grande contribuição, uma vez que, em geral, conjuntos de dados com maiores quantidades de dados rotulados tendem a treinar modelos de classificação com maiores F-Scores.

A análise de sentimentos levando em consideração a análise de *emoticons* também é outro ponto que poderia ser trabalhado futuramente, uma vez que *emoticons* podem ser úteis para classificação de emoções e podem contribuir para uma elevação na taxa de acertos das predições. Da mesma forma, o emprego da categoria *topic-based methods* na abordagem de aprendizado semi-supervisionado seria uma contribuição relevante para o estudo.

A utilização de modelos de aprendizado de máquina não-supervisionados nos experimentos também seria uma boa contribuição ao projeto, visto a dificuldade e complexidade envolvida na rotulação de conjuntos de dados. Assim como, adicionar modelos de aprendizagem profunda em nossos experimentos, pois estes se mostraram relevantes em trabalhos relacionados à análise de sentimentos.

Por fim, para análises voltadas à saúde mental na Universidade, outro ponto que poderia ser levado em consideração seria a coleta de *tweets* de acordo com o período do semestre, para que possa ser realizada uma comparação entre sentimentos expressos no começo e no fim de cada semestre universitário. Adicionalmente, uma análise mais detalhada das emoções predominantes encontradas nos experimentos pode gerar discussões relevantes para melhorias na saúde mental de estudantes e docentes da Universidade.

Referências

[1] Gerber, Matthew S: *Predicting crime using twitter and kernel density estimation.* Decision Support Systems, 61:115–125, 2014. 1

[2] Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson e David Lazer: *Fake news on twitter during the 2016 us presidential election.* Science, 363(6425):374–378, 2019. 1

[3] Aguiar, Erikson Júlio de, Bruno S Faiçal, Jó Ueyama, Glauco Carlos Silva e André Menolli: *Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação.* Em *Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*. SBC, 2018. 1, 2, 16, 18

[4] Silva, Nadia Felix F Da, Luiz FS Coletta e Eduardo R Hruschka: *A survey and comparative study of tweet sentiment analysis via semi-supervised learning.* ACM Computing Surveys (CSUR), 49(1):1–26, 2016. 2, 9, 10, 16, 17, 25, 26, 27

[5] Blumberg, Robert e Shaku Atre: *The problem with unstructured data.* Dm Review, 13(42-49):62, 2003. 5

[6] Aggarwal, Charu C e ChengXiang Zhai: *Mining text data.* Springer Science & Business Media, 2012. 6

[7] Loh, Stanley, Leandro Krug Wives e Antônio Severo Frainer: *Uma abordagem para a busca contextual de documentos na internet.* Revsita de Informática Teórica e Aplicada (RITA), 4(2):79–92, 1997. 6

[8] Saif, Hassan, Miriam Fernandez, Yulan He e Harith Alani: *On stopwords, filtering and data sparsity for sentiment analysis of twitter.* Em *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), may 2014, ISBN 978-2-9517408-8-4. 6

[9] Manning, Christopher D., Prabhakar Raghavan e Hinrich Schütze: *Introduction to Information Retrieval.* Cambridge University Press, USA, 2008, ISBN 0521865719. 7

[10] Rajput, Quratulain, Sajjad Haider e Sayeed Ghani: *Lexicon-based sentiment analysis of teachers' evaluation.* Applied Computational Intelligence and Soft Computing, 2016:1–12, janeiro 2016. 7

[11] Wolny, Wiesław: *Sentiment analysis of twitter data using emoticons and emoji ideograms*. Studia Ekonomiczne, 296:163–171, 2016. 7

[12] Goldberg, Yoav e Graeme Hirst: *Neural network methods in natural language processing. morgan & claypool publishers(2017)*. 9781627052986 (zitiert auf Seite 69). 7

[13] Qaiser, Shahzad e Ramsha Ali: *Text mining: use of tf-idf to examine the relevance of words to documents*. International Journal of Computer Applications, 181(1):25–29, 2018. 8

[14] Mitchell, Thomas M.: *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1ª edição, 1997, ISBN 0070428077, 9780070428072. 9

[15] Monard, Maria Carolina e José Augusto Baranauskas: *Conceitos sobre aprendizado de máquina*. Sistemas inteligentes-Fundamentos e aplicações, 1(1):32, 2003. 9

[16] Balcan, Maria Florina e Avrim Blum: *A discriminative model for semi-supervised learning*. Journal of the ACM (JACM), 57(3):19, 2010. 10

[17] Zhu, Xiaojin e Andrew B Goldberg: *Introduction to semi-supervised learning*. Synthesis lectures on artificial intelligence and machine learning, 3(1):1–130, 2009. 10

[18] Smola, Alex J e Bernhard Schölkopf: *A tutorial on support vector regression*. Statistics and computing, 14(3):199–222, 2004. 11

[19] Smola, Alexander J, Peter J Bartlett, Dale Schuurmans e Bernhard Schölkopf: *Advances in large margin classifiers*. MIT press, 2000. 11

[20] Rish, Irina *et al.*: *An empirical study of the naive bayes classifier*. Em *International Joint Conference on Artificial Intelligence (IJCAI) - 2001 workshop on empirical methods in artificial intelligence*, volume 3, páginas 41–46, 2001. 11, 12

[21] McCallum, Andrew e Kamal Nigam: *A comparison of event models for naive bayes text classification*. Em *Association for the Advancement of Artificial Intelligence (AAAI)-98 workshop on learning for text categorization*, volume 752, páginas 41–48. Citeseer, 1998. 12, 25

[22] Peterson, L. E.: *K-nearest neighbor*. Scholarpedia, 4(2):1883, 2009. revision #137311. 12

[23] Zhu, Xiaojin e Zoubin Ghahramani: *Learning from labeled and unlabeled data with label propagation (technical report cmu-cald-02-107)*. Carnegie Mel-lon University, 2002. 13

[24] Johnson, Christopher, Parul Shukla e Shilpa Shukla: *On classifying the political sentiment of tweets*. Cs. utexas. edu, 2012. 13

[25] Ting, Kai Ming: *Confusion Matrix*, páginas 260–260. Springer US, Boston, MA, 2017, ISBN 978-1-4899-7687-1. 14

[26] Ting, Kai Ming: *Precision*, páginas 990–990. Springer US, Boston, MA, 2017, ISBN 978-1-4899-7687-1. 14

[27] Li, Gang e Fei Liu: *A clustering-based approach on sentiment analysis.* Em *2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering*, páginas 331–337. IEEE, 2010. 17, 25

[28] Hong, Liangjie e Brian D Davison: *Empirical study of topic modeling in twitter.* Em *Proceedings of the first workshop on social media analytics*, páginas 80–88. acm, 2010. 18, 25

[29] Bíró, István, Jácint Szabó e András A Benczúr: *Latent dirichlet allocation in web spam filtering.* Em *Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, páginas 29–32, 2008. 18

[30] Dias, Marcelo e Karin Becker: *Detecção semi-supervisionada de posicionamento em tweets baseada em regras de sentimento.* Em *SBBD*, páginas 40–51, 2016. 18

[31] Rosenthal, Sara, Alan Ritter, Preslav Nakov e Veselin Stoyanov: *SemEval-2014 task 9: Sentiment analysis in twitter.* Em *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, páginas 73–80, Dublin, Ireland, agosto 2014. Association for Computational Linguistics. 20, 45

[32] Navega, Sergio: *Princípios essenciais do data mining.* Anais do Infoimagem, 2002. 24

[33] Nakov, Preslav, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani e Veselin Stoyanov: *Semeval-2016 task 4: Sentiment analysis in twitter.* arXiv preprint arXiv:1912.01973, 2019. 30